Detection of genomic signatures of selection in roe deer and reindeer populations

DE-JONG, MENNO,JEROEN

Detection of genomic signatures of selection in roe deer and reindeer populations

Menno Jeroen de Jong

This thesis is submitted in candidature for the degree of

Doctor of Philosophy

Biosciences department

Durham University

2019

Detection of genomic signatures of selection in roe deer and reindeer populations

Menno Jeroen de Jong

This thesis is submitted in candidature for the degree of

Doctor of Philosophy

Biosciences department
Durham University
2019

Abstract

In this thesis I present the outcomes of genetic analyses of several reindeer and roe deer datasets, using two types of data: single nucleotide polymorphism (SNP) data and whole genome sequencing data. I assess the population structure, genetic diversity and demographic history of the study populations and study species, but the main focus is on selection analyses: the detection of genetic signals of selection.

In Chapter 2 I present SNP data analysis outcomes which are suggestive of a shared positive selection event in two reindeer founder populations on the South Atlantic island South Georgia. This finding therefore possibly provides empirical evidence that positive selection can overcome drift in heavily bottlenecked founder populations, and can be detected despite elevated background neutral variation. In addition, I report a new selection scan called Genome Wide Differentiation Scan (GWDS).

In Chapter 3 I infer from a SNP dataset that the effective population size of the native UK roe deer population has numbered several thousand individuals throughout the Holocene. The dataset suggests that neither drift nor positive selection has caused fixed differences between the UK population and the European mainland population, despite a split time of ~1500 generations.

In Chapter 4 I investigate the demographic and evolutionary history of the extant roe deer sister species: the European roe deer (*C. capreolus*) and the Siberian roe deer (*C. pygargus*). Whole genome sequences analyses suggest that the two species split maximum 1.6Mya and show pronounced differences in terms of genetic diversity and effective population sizes (Ne). In the species with lower genetic diversity and lower historical Ne, *C. capreolus*, I find higher proportions of lineage specific amino acid substitutions. This negative relationship between Ne and number of non-synonymous substitutions is suggestive of relaxation of purifying selection, but alternative explanations (such as episodes of positive selection and data artifacts resulting from differences in genome quality) can not be excluded.

In Chapter 5 I discuss the results presented in this thesis in the light of the neutral theory of molecular evolution.

Table of contents

Declaration

The material contained in this thesis has not previously been submitted for a degree at the University of Durham or any other university. The research reported within this thesis has been conducted by the author unless otherwise stated.

Acknowledgements

Countless persons have contributed to my PhD projects, and many of them I don't know personally or even by name. Early on in my PhD I decided that the acknowledgements in my PhD-thesis should start with expressing gratitute to everyone who is or has been taking the effort to provide online bioinformatics support. It is difficult to imagine working as a bio-informatician without the possibility to google for awk one-liners, for R function options or for solutions to unfamiliar error messages. Thanks also to the developers of the numerous software packages I have used during my PhD, especially to those who took the time and effort to write a detailed manual.

The first non-anonymous person I should thank is my professor Rus Hoelzel, first and foremost for his very close supervision, but also for providing me this PhD-studentship opportunity. I know that you simply selected the student which you thought was most suitable for the project, but for me it was a life saver. When I started this PhD I had been applying for almost two years, having to get by with temporary employments and in need of career satisfication. My bioinformatics skills were limited. I didn't know how to run a simple for-loop or how to view a bam file. Much has changed in the 4.5 years of my PhD. I have learnt almost every day and developed the skills needed to make contributions to a research field which fascinates me.

Another person I am greatly indebted to, literally even, is the late George Kenneth Whitehead, a deer enthusiast who left his fortune to Durham University for research on deer species. Without his financial contribution I wouldn't have had the opportunity to do this PhD. I should also thank the members of the British Deer Society, because they co-funded my PhD.

My acknowledgments wouldn't be complete without thanking the persons who helped me to acquire the samples and/or datasets analysed in this thesis. I am grateful to Zhipeng Li and Wen Wang for providing the whole genome assembly of the *C. pygargus* genome, to Erwan Quemere for providing ddRADseq data on a

French roe deer population, and to Regina Kropatsch for providing the sequencing reads of the *C. capreolus* genome. I furthermore thank Hugh Rose and Annette Kohnen, as well as associated stalkers, for their indispensable efforts to provide roe deer samples from Wurttemberg (Germany) and East Anglia (UK).

I thank Karis for getting me started in the lab, and Michelle for helping me getting to grisps with the lab work and for guiding my first steps into the world of bioinformatics. My brother, Joost, introduced me to SNP data analyses and sent me useful R scripts, which laid the foundation of my R package 'SambaR'.

My parents helped me move to the UK and have come to see me every year. Thank you for your ever-lasting support. Also thanks to all other Dutch friends and family members who have visited me.

During my time in Durham I have been lucky enough to meet many new friends from all over the world and they have made my life so much richer and so much more fun. One person has been particularly nice to me, and I am very grateful that one day I was presented the opportunity to get into contact with her. Thank you, Fearn, for being such a lovely girlfriend. I couldn't wish for more.

Over the last 4.5 years I have been sitting almost every day on the same chair in the corner of lab 3, staring at my computer screen. But for me it doesn't feel like that at all. In my memories I picture myself together with friends, exploring the North East, the Yorkshire Moors and the Scottish Highlands, either by car, by foot, or by bike. I will take these wonderful memories back home, and I hope they will last for long.

Durham, 26-11-2019

Publication list

Peer-reviewed publications reporting the analyses and results presented in this PhD-thesis:

Chapter 2:

De Jong, M.J., Lovatt, F., Hoelzel, A.R., *Detecting genetic signals of selection in heavily bottlenecked founder reindeer populations by comparing parallel founder events*
Under review by Molecular Ecology
Author contributions: ARH conceived the study and MdJ & ARH wrote the paper. MdJ undertook data, simulation and lab analyses, and developed the selection scan GWDS. FL provided field work and some of the DNA extractions.

Chapter 2 and 3:

De Jong, M.J., De Jong, J.F., Hoelzel, A.R., Janke, A., *SambaR: an R package for fast, easy and reproducible population-genetic analyses of biallelic SNP datasets*
Published as preprint at bioRxiv and under review by Molecular Ecology Resources
Author contributions: MdJ and JdJ developed the software. MdJ wrote the paper and the manual, and developed the selection scan GWDS, the BPA-test and the dc-score. ARH and AJ provided funded, feedback/advice, and input to the writing

Chapter 3 and 4:

De Jong, M.J., Li., Z., Qin, Y., Quemere, E., Baker, K., Wang, W. 2020. *Demography and adaptation promoting evolutionary transitions in a mammalian genus that diversified during the Pleistocene*
Published in Molecular Ecology
Author contributions: ARH conceived the study and MdJ & ARH wrote the paper. MdJ undertook data and lab analyses. EQ provided a subset of the RADseq data. ZL, YQ and WW generated the *C. pygargus* genome assembly and annotation.

Chapter 1

General Introduction

Overview

This thesis investigates the genetic divergence of populations and species over time. More specifically, it investigates to what extent genetic divergence is driven by genetic drift and to what extent by natural selection.

Although the thesis does include some modelling and simulations, it is centred around the analysis of empirical datasets. These datasets comprise whole genome sequences and single nucleotide polymorphism (SNP) datasets derived from wild populations of deer, more precisely reindeer (*Rangifer tarandus*), the western aka European roe deer (*Capreolus capreolus*), and the eastern aka Siberian roe deer (*Capreolus pygargus*). The thesis is built around three such datasets, all of which have in common that they allow for comparisons between sister taxa (i.e. closely related populations or species that together with their ancestral population/species constitute a monophyletic taxon).

The main difference between the three datasets is the age of the sister taxa, also known as the time to the most recent common ancestor (TMCRA). The TMRCA of the sister populations in the three datasets ranges over orders of magnitude, from 10^2 years to 10^6 years. This enables an exploration of genetic divergence – and the contribution of natural selection – on various time scales and consequently of various evolutionary stages of cladogenesis, from incipient population differentiation to post-speciation differentiation.

Research on the role of natural selection in driving the genetic divergence of populations and species is needed to settle a long-standing debate within the reseach field of evolutionary biology. According to the neutral theory of molecular evolution, first posed in 1968, most genetic differences between populations and species within protein-coding DNA are due to neutral substitutions instead of adaptively driven substitutions (Kimura, 1991). Although the theory exists for half a century, evolutionary biologists are still divided over the validity of this claim (Jensen et al., 2019; Kern and Hahn, 2018). In this first Chapter of my thesis, the

general introduction, I will discuss the neutral theory in detail and explain how studies such as the ones presented in this thesis, can contribute to the discussion.

Drift or selection?

Whenever a population splits into geographically separated sister populations, these sister populations will start to diverge both genetically and phenotypically. The genetic differences, which are the focus of this thesis, can range in size from whole chromosome duplications and rearrangements to single nucleotide variations (SNVs).

The origin of new genetic variation – the first appearance of a new genetic variant (i.e. allele) in the population or of new combinations of genetic variants – depends on stochastic processes solely: mutations and recombinations. The fate of the differences – whether or not they spread throughout the population – is governed by an interplay of two processes, one of which is deterministic and the other stochastic: natural selection (Darwin and Wallace, 1858) and genetic drift (Wright, 1931) respectively. Genetic drift is defined as allele frequency change through random sampling. Random sampling refers to random survival and reproduction of individuals. Natural selection is the opposite of random sampling, and occurs when certain individuals have a higher survival and reproduction probability due to a certain beneficial phenotypic trait. As populations are finite by nature, selection never works completely in isolation from genetic drift, but the bigger the number of breeders, the smaller the influence of drift (Hartl and Clark, 1997).

The advent of sequencing techniques has allowed us to characterize the structure and extent of genetic variation among populations and species in detail. The current challenge for molecular and evolutionary biologists is to identify which genetic variation is functional as well as which genetic variation is non-neutral. Questions which can be posed are: 1.) Which genetic differences cause the observed phenotypic differences between populations and/or species?; and 2.) How, why and when did these differences establish? (Varki and Altheide, 2005).

Whereas the findings of functional genomics clarify which parts of the genome affect the phenotype, it is the aim of selection analysis to infer which genetic changes in those regions are driven by selection and which by genetic drift. The

investigations of my PhD thesis fall within this latter research field of selection analysis (and hence not in the field of functional genomics).

Bridging the gap

The questions about the genetic divergence of populations and species over time, and the relative importance of selection and drift in driving this divergence process, touch upon a deep gap in our understanding of the inner workings of evolution: the translation from microevolution to macroevolutionary phenomena (Reznick and Ricklefs, 2009; Uyeda et al. 2011; Pennell et al. 2013). Although macroevolution entails both the diversification (cladogenetic speciation) and succession (anagenetic speciation and evolutionary innovations) of life on Earth, here I will focus mostly on cladogenetic speciation, because this is what my datasets allowed me to investigate.

Despite the book's title 'On the Origin of Species', Darwin did not solve the riddle of speciation (Mayr, 1999) but rather the riddle of adaptation. Rather than describing the entire speciation process, he described a fundamental, repetitive step of the process, a step which can be defined as the fixation of genetic mutations by selection. This left open many follow-up question, including the question how many of these adaptive steps are needed to progress through speciation (Via, 2009).

We know now that the process of speciation contains at least one more building block: the fixation of genetic mutations by genetic drift. The speciation process can therefore be envisioned as a cumulative process of both selective and neutral substitution events, with speciation as ultimate outcome. Hence the question refines to: how many steps make up the speciation process, and how many of those steps are adaptive steps and how many are neutral steps? The addition of a second building block in the process of speciation furthermore opens up the possibilities for different modes of speciation, characterized by different proportions of neutral and adaptive steps.

According to Mayr's biological species concept (BSC, Mayr, 1999), the speciation process is completed once individuals from both sister populations can no longer interbreed. A common implicit assumption behind speciation models is that the establishment of reproductive isolation, and hence the formation of a species, is a 'by-product' of neutral and adaptive genetic and phenotypic divergence (Schluter, 2001; Sobel et al., 2010). Natural selection does not directly favour genetic

incompatabilities (Seehausen et al., 2014) or phenotypic traits that prohibit gene flow. Instead, reproductive isolation is the indirect consequence of the genetic and phenotypic divergence of sister populations over time (Via, 2009), possibly catalyzed by sexual selection (Wellenreuther and Sánchez-Guillén, 2016). The alternative model, according to which selection disfavours hybrids, is called speciation by reinforcement (Hoskin et al., 2005).

A debate has been ongoing for decades about the exact nature of the adaptations driving speciation and species replacement. Natural selection is an umbrella term for a myriad of selective pressures, and can be categorized in distinct classes such as biotic and abiotic driven selection, interspecies interactions and intraspecies competition selection, resource and predator driven selection, and intrasex and intersex sexual selection. Whereas originally interspecies competition was and still is regarded as a main driver of macroevolution, paleontologists have argued in favour of abiotic factors rather than biotic factors being the main drivers (Benton, 2009). Another open question is whether natural selection mostly works on new mutations in stable environments (mutation driven selection), or on standing variation following environmental change or migration into new environments (Van Valen, 1963; Barrett and Schluter, 2008). In this thesis I will ignore all these subcategories of natural selection, and discriminate between neutral events and selective events without considering or questioning the exact nature of the selective events.

Speciation modes

Our limited understanding of macroevolution illustrates the existence of boundaries of empirical research. Whereas the process of adaptation lays within the realm of direct observation and/or experimental manipulation, the process of speciation lays outside this realm. Although there is some evidence for the generation of reproductive barriers within ecological time frames (Hendry et al., 2007; Lamichhaney et al., 2018; McKinnon et al., 2004; Montesinos et al., 2012), there is reason to believe that speciation – apart from polyploidy speciation – typically requires timespans of 10^5-10^6 years (Avise, 2000; Avise et al., 1998; Curnoe et al., 2006; Lister, 2004). In addition, whereas microevolutionary events follow relatively few and simple rules, the process of macroevolution has many unknowns,

prohibiting insights from mathematical and modelling studies. As a result, the study of macroevolution is a historic science (Kemp, 2007), relying on incomplete evidence.

Inferences about the process of speciation can be drawn from biogeographical data. During the 19th century naturalists such as Moritz Wagner and Alfred Wallace observed that sister species often occur in adjacent regions which are separated by a geographical barrier such as a river or a mountain range. From these observations Wagner deduced his natural law of allopatric speciation. He wrote: 'The formation of an incipient species can succeed in nature only when some individuals can cross the previous borders of their range and segregate themselves for a long period from other members of their species.' (Schilthuizen, 2002)

The universality of allopatric speciation has been questioned by Darwin (Schilthuizen, 2002) and many other evolutionary biologists since, who argued in favour of either parapatric (Endler, 1977) or sympatric speciation (Schilthuizen, 2002). A fourth demographic mode of speciation, which can be regarded as a subcategory of allopatric speciation, was suggested by Mayr. Mayr observed that islands hold a disproportional number of endemic species, which led him to induce the peripatric or bottleneck speciation model, according to which founder events facilitate speciation (Mayr, 1999; Templeton, 2008). Metastudies seem to point to allopatric speciation as the main geographic mode of speciation (Barraclough and Vogler, 2000), but have also provided some evidence for alternative modes, including bottleneck speciation (Barraclough and Vogler, 2000; Vrba and DeGusta, 2004).

Different geographical modes of speciation might involve different relative contributions of drift and selection. Drift is presumably especially dominant in bottleneck speciation. Mayr put forward his 'genetic reconstruction'-hypothesis which states that founder events facilitate speciation through stochastic factors (Mayr, 1954). He argued that by randomly altering allele frequencies, bottleneck events affect epistatic effects (i.e. gene-gene interactions), resulting in a genetic and phenotypic 'revolution' (Barton and Charlesworth, 1984). Hampton Carson expanded the bottleneck speciation model to the founder-flush speciation model by suggesting an additional explanation for bottleneck speciation. He hypothesized that during the population expansion phase – the time window spanning from the

founder bottleneck to the moment the founder population reaches its carrying capacity – purifying selection is relaxed, facilitating the fixation of slightly deleterious alleles (Templeton, 2008). In short, both Mayr and Carson attributed bottleneck speciation to genetic drift, not to natural selection.

Mayr also hypothesized a mechanism behind the presumably more prevalent mode of allopatric speciation. Again, he considered an important role for drift. Like Wagner, he noted that extant sister species often occur in geographically isolated habitats. These habitats were, although geographically isolated, often environmentally similar. This observation potentially questions the importance of adaptation in driving speciation. It can be argued that sister populations which occur in similar environments can evolve in different directions, because mutations arise randomly and populations therefore can adapt in different ways to similar environmental conditions (mutation-order speciation – see definition below). But an alternative explanation is that speciation can occur without the help of natural selection, through drift only.

Speciation modes can thus be defined not only based on geographical distribution of the incipient sister species, but also on the driving forces behind the divergence process. As such, a distinction can be made between two hypothetical extremes: ecological speciation and neutral speciation (aka non-ecological speciation) (Baptestini et al., 2013; Gittenberger, 1991; Reaney et al., 2018; Rundell and Price, 2009; Stuessy et al., 2006).

In the neutral speciation model populations diverge and eventually speciate through random fixation of mutations rather than selective driven fixation. In this model, the role of natural selection is downgraded from main driver of change to that of catalyst. Geographical separation in itself is sufficient for populations to diverge, and, given enough time, to result in speciation. Natural selection, in particular sexual selection, can speed up the process and cause reproductive isolation, but is not strictly needed (Czekanski-Moir and Rundell, 2019; Janecka et al., 2012; Wellenreuther and Sánchez-Guillén, 2016).

In the ecological speciation model populations diverge and eventually speciate through selection driven fixation of mutations (Schluter, 2009). Ecological speciation can be grouped in two broad categories: divergence of sister populations adapting to contrasting environments (the narrow definition of ecological

speciation), and divergence of sister populations adapting in different ways to similar environments, termed mutation-order speciation, non-ecological speciation (Schluter, 2009), or uniform selection speciation (Sobel et al., 2010). This latter scenario might be especially likely if adaptation is many driven by abiotic selection pressures, such as intraspecies and interspecies competition, which are supposedly less dependent on geographical distribution as adaptations to abiotic selection pressures.

Neutral speciation and ecological speciation are potentially theoretical constructs which do not exist in nature. Rather, they might represent the opposite ends of a speciation spectrum in which neutral forces and selective pressures contribute in varying relative strengths to species divergence, resulting in different proportions of selective driven substitutions (i.e. different estimates of alpha, discussed below). As environments are highly multi-dimensional, sister populations are presumably rarely exposed to identical selection pressures, meaning that the divergence of populations is rarely driven by drift alone, and that some substitutions will be pushed, perhaps only marginally, by selection. Likewise, even when selection is a main driver of population divergence, it still holds that parts of the genome are non-functional nor tightly linked to adaptive functional regions, and therefore that a certain proportion of substitutions will be driven by drift.

The neutral theory of molecular evolution

In the 1960's the development of protein sequencing methods (Chadarevian, 1999) and gel electrophoresis (Smithies, 2012) enabled direct inference about genomic evolution, rather than from indirect lines of evidence such as biogeography. The new insights inspired a theory about genomic evolution which has never been free from controversy but yet has remained the dominant theory to the present day: the neutral theory of molecular evolution.

The neutral theory was nearly simultaneously proposed in two papers, one published in *Nature* (Kimura, 1968), the other shortly after in *Science* (King and Jukes, 1969). The rather uninspiring title of Kimura's paper, 'Evolutionary rate at the molecular level', obscured its main and controversial selling point, namely that most substitutions are driven by drift and not by selection. The King and Jukes (1969) paper, in contrast, was provocatively titled 'Non-Darwinian Evolution'. It

was meant to provoke, and so it did (King, 1983). The controversy started even before publication, during the review process. The King and Jukes (1969) paper was accepted only after rebuttal. The reasons for the initial rejection were contradictory. Jack King, one of the two authors, later recalled: 'One referee said that we had merely set up and demolished a straw man and that the idea was obviously true and therefore trivial. The other said the idea was obviously false.' (King, 1983)

The short abstracts of both papers capture the essence of the neutral theory. The abstract of Kimura (1968) reads: 'Calculating the rate of evolution in terms of nucleotide substitutions seems to give a value so high that many of the mutations involved must be neutral ones.' The abstract of King and Jukes (1969) was even shorter: 'Most evolutionary change in proteins may be due to neutral mutations and genetic drift.' This is the neutral theory stripped down to its bare essence: a single proposition, stating that most nucleotide substitutions are neutral, not adaptive.

Table 1.1 The (nearly) neutral theory of molecular evolution

class	*Mutation proportion*	*Fixation probability*	*Substitution proportion*
beneficial	very low	high	low
neutral	*K68*: high *KJ69:* low (≤10%) *KO71*: low *O73*: low	1/(2·Ne)	*K68*: high *KJ69*: high *KO71*: high *O73*: low if |s| << 1/(2·Ne), high if |s| >> 1/(2·Ne)
deleterious	*K68*: low *KJ69*: high *KO71*: high *O73*: high	*K68*: very low *KJ69*: very low *KO71*: very low *O73*: 1/(2·Ne) if |s| << 1/(2·Ne), very low if |s| >> 1/(2·Ne)	*K68*: low *KJ69*: low *KO71*: low *O73*: high if |s| << 1/(2·Ne), very low if |s| >> 1/(2·Ne)

K68: Kimura, 1968; *KJ69:* King and Jukes, 1969; *KO71*: Kimura and Ohta, 1971; *O73*: Ohta, 1973

But the neutral theory also provides an explanation, a mechanism, for the prevalence of neutral substitutions. This explanation rests upon the concepts of mutation rate and fixation probability, and how these factors differ among three classes of mutations: beneficial, neutral and deleterious mutations (Table 1.1). The theory holds that deleterious mutations occur frequently but that only a very small proportion manages to escape purifying selection and to reach fixation. Beneficial mutations behave in the opposite way: they occur rarely, but if they do, they ordinarily reach fixation, due to the workings of positive selection. Neutral alleles

have a winning intermediate strategy: their fixation probability is relatively low compared to beneficial mutations but high compared to deleterious mutations, and since they occur frequently, this still adds up to a high number. As I will discuss below, refinements of the neutral theory have led to slightly different versions of the neutral theory (Table 1.1). The main conclusion remains however unchanged: The net effect of the two factors, mutation rate and fixation probability per mutation class, is that only a small proportion of all substitutions are adaptive (Table 1.1).

The major argument provided by Kimura (1968) in favour of the proposition that most substitutions are neutral, was that the observed substitution rate in nature was so high it could not be explained by selection. Kimura (1968) furthermore claimed that the observed level of genetic variation within populations also agreed with the proposition. He would elaborate this argument in a second paper (Kimura and Ohta, 1971). Whereas Kimura supported the proposition using considerations from the field of theoretical population genetics, King and Jukes (1969) came up with a list of arguments from the field of molecular biology. As a result, the thinking about the dominance of neutral mutations and random drift expanded to a coherent set of ideas, worthy of the label theory – the 'neutral mutation-random drift theory', as Kimura and Ohta (1971) originally called it. The core of this theory, the main proposition and the underlying mechanism, became framed by a set of testable predictions which were deduced from either the proposition or the underlying mechanism, and which could be tested against the growing amount of available sequence data. In the words of Kimura and Ohta (1971): 'The neutral mutation-random drift theory allows us to make a number of definite quantitative as well as qualitative predictions by which the theory can be tested. We hope that through this process we will be able to gain deeper understanding of the mechanism of evolution at the molecular level and will be emancipated from a naïve pan-selectionism.'

The growing complexity of the theory cultivated several misunderstandings which confound the debate about the theory. The essence of the neutral theory is for example not that many mutations are neutral or deleterious – few selectionists would argue with this (Kern and Hahn, 2018). The essence is that the majority of substitutions are the result of stochastic fixation of these neutral mutations, rather than the result of selective driven fixation of adaptive alleles (Table 1.1).

The neutral theory has been interpreted to mean that differences between species are caused by these non-adaptive substitutions (Kern and Hahn, 2018), but this is not what the founders of the theory believed. The neutral theory of molecular evolution is a theory about genomic evolution, not about phenotypic evolution. Although the neutral theory holds that most substitutions are neutral (or slightly deleterious), it does not rule out the possibility that the phenotypic differences observed between species is caused by the minority of adaptive substitutions. This decoupling of genomic and phenotypic differences was stressed by King and Jukes (1969). From their introductory remarks it is evident that even though they argued that most nucleotide substitutions in proteins are neutral, they believed that most species differences at the phenotypic level are adaptive. 'Evolutionary change at the morphological, functional and behavioral levels,' they wrote, 'results from the process of natural selection, operating though adaptive changes in DNA. It does not necessarily follow that all or most evolutionary change in DNA is due to the action of Darwinian natural selection.'

Another misunderstanding is that the neutral theory partly rests upon the vast majority of the genome being non-coding (Kern and Hahn, 2018; Jensen et al., 2019). This was however not part of the original argumentation. Although King and Jukes (1969) did discuss the presence of non-coding DNA, they did so in a different context, as will be discussed below. The neutral theory was developed in a time that actual sequence data was sparse and limited to proteins. The theory was developed to explain observed patterns in these data sets. As a consequence, the original arguments for the neutral theory pertained to proteins, not to full genomes. This is reflected in the abstract of Kimura and Ohta (1971), which reads: 'It is proposed that random genetic drift of neutral mutations in finite populations can account for protein polymorphisms.' And similarly, in the abstract of King and Jukes (1969), which reads: 'Most evolutionary change in proteins may be due to neutral mutations and genetic drift.' It is therefore a fallacy to argue that most substitutions are neutral because the majority of the genome is non-coding.

Because it is my personal belief that a theory is best understood by knowing the history of the theory, I will discuss the theory by means of a historical account. I will first lay out the original reasoning which led Kimura (1968), King and Jukes

(1969) and Kimura and Ohta (1971) to propose the theory, and afterwards describe the counterarguments.

Haldane's dilemma

In his landmark 1968 paper, which laid the basis of the neutral theory, Kimura made two main statements. First, he calculated that amino-acid and nucleotide substitution rates occurred in nature in much higher rates than previously thought or even held possible. Second, he argued that this rate could not exist if most mutations were adaptive. This led him to reject this hypothesis and instead formulate an alternative hypothesis that was consistent with the observed substitution rate – namely the hypothesis that most substitutions are neutral.

The dilemma addressed by Kimura is now known as Haldane's dilemma, named after John Haldane, who in 1957 had published an influential paper titled 'The cost of natural selection'. In here Haldane had put forward 'the fairly obvious statement' that since adaptation comes with the cost of additional mortality, the reproductive capacity of organisms puts an upper limit to the rate of evolution (Haldane, 1957) and therefore to the amount of genetic differences between species. The exact nature of Haldane's calculations, and whether they led to the right conclusions, go too much into detail to be discussed here. It suffices to say that Haldane concluded that for animals with relatively low reproductive capacities (i.e. low number of offspring per adult per generation), such as most vertebrates, the upper rate of molecular evolution was limited to 1 nucleotide substitution per 300 generations (Haldane, 1957).

Haldane's upper limit stood in sharp contrast to insights obtained from the new data on genetic divergence between species, acquired through protein gel electrophoresis. Kimura calculated, based on at the time available data of three proteins (haemoglobin, cytochome c and triosephosphate dehydrogenase) that the average interval time between two subsequent nucleotide mutations was 1.8 years, much lower than the minimum interval time of 300 years calculated by Haldane (Kimura, 1968).

To reconcile the observed high proportion of genetic differences with Haldane's calculations, Kimura proposed his 'mutation-random drift theory', now better known as the neutral theory of molecular evolution (Kimura, 1968, 1991).

The theoretical problem presented by Haldane's calculations, Kimura argued, dissolved if most substitutions were driven by drift rather than by selection. His assumption was that drift did not involve differential mortality, and therefore would put less strain (i.e. lower death toll) on a population. Kimura: 'For a nearly neutral mutation the substitutional load can be very low and there will no limit to the rate of gene substitution in evolution.' (Kimura, 1968).

Kimura, whose many contributions to science included mathematical work on allele fixation probabilities (Kimura, 1962; Kimura and Ohta, 1969), showed that the substitution rates of neutral alleles equals the mutation rate (Kimura, 1968). Therefore, an estimate of the neutral substitution rate could be obtained by simply multiplying the mutation rate by the genome size. This led to the conclusion that neutral substitutions must occur very frequently, close to Kimura's estimate of 1 substitution every 1.8 year. (Or in fact more frequent even. For example, given a genome size of 3 Gb and a mutation rate of $2.2*10^{-9}$ per year (Kumar and Subramanian, 2002), the substitution rate is 6.6 substitutions per year).

The implication was that observed substitution rates made for a closer match with neutral expectations than with theories based on selection. Kimura therefore argued that, contrary to the perception of the time, genetic drift was a dominant force in driving genomic evolution. 'The significance of random genetic drift has been deprecated during the past decade', Kimura wrote towards the end of his paper. 'This attitude has been influenced by the opinion that almost no mutations are neutral, and also that the number of individuals forming a species is usually so large that random sampling of gametes should be negligible in determining the course of evolution, except possibly through the founder principle.' (Kimura, 1968) It was time to rethink the role of drift, Kimura stated. 'To emphasize the founder principle but deny the importance of random genetic drift due to finite population number is, in my opinion, rather similar to assuming a great flood to explain the formation of deep valleys but rejecting a gradual but long lasting process of erosion by water as insufficient to produce such a result.' (Kimura, 1968)

Substitutional load

A key argument in Kimura's 1968 paper is that fixation of alleles through drift does not involve additional mortality caused by selection. Although this assumption

seems self-explanatory and not in need of further evidence, Kimura did back up this argument mathematically by presenting a new formula he had derived in the previous months. Kimura promised to publish the derivation of this formula elsewhere, and he did so one year later, in the journal *Heredity* (Kimura and Maruyama, 1969).

Apart from a previously published formula on fixation probability (Kimura, 1957), it is the only formula in his 1968 landmark paper. The explanatory variables were the selection coefficient and effective population size. The dependent variable was the substitutional load, which Kimura defined as the temporary lowering of the mean population fitness during the substitution process. Between brackets Kimura mentioned that this substitutional load was his 'terminology' for Haldane's 'selection intensity' (Haldane, 1957), the proportion of deaths which are selective. The two concepts are indeed closely related: a fitness difference (substitutional load) quantifies the proportional difference in surviving offspring (selection intensity).

Kimura's formula showed the obvious, namely that the substitutional load of neutral alleles equals zero. However, importantly, the formula also showed that even alleles which are not completely neutral, can still be effectively neutral, depending on the population size. The magnitude of drift is inversely related to the effective population size. The lower the selection coefficient in comparison to the effective population size, the smaller the substitutional load. For alleles for which the selection coefficient was smaller than the inverse of the effective population size, the substitutional load converged to zero. Kimura referred to these nearly neutral alleles as 'the special case of $2 \cdot Ne \cdot s << 1$'. He concluded: 'For a nearly neutral mutation the substitutional load can be very low and there will no limit to the rate of gene substitution in evolution.' (Kimura, 1968).

Although this interplay between drift, selection and the population size is nowadays part of mainstream thought, at the time this was a novel insight, even to Kimura. Investigation of Kimura's earlier work on the substitutional load might help us to understand better what led Kimura to understand the importance of drift.

In 1960, eight years before his landmark 1968 paper, Kimura had published his first paper on the subject (Kimura, 1960). In this paper Kimura set out to mathematically derive the optimum mutation rate. He noted that most mutations

are deleterious, and therefore that the occurrence of mutant alleles in a population, maintained in a mutation-drift equilibrium, generally decreases individual fitnesses. On the other hand, if a population was devoid of any genetic variation, then this population did not contain standing genetic variation needed for adaptation when confronted with environmental change. In Kimura's own words: 'The higher the mutation rate, the more the reproductive potential of a species will be impaired. Yet, without heritable variation, adaptive evolution by natural selection will be impossible. If gene mutation ceases to occur, the store of genetic variability of a species will soon be depleted; and when environmental conditions change, the species will no longer be able to readjust itself to the new environment.' (Kimura, 1960)

Kimura reasoned that the trade-off between these conflicting costs should have resulted in an optimum mutation rate. He wrote: 'These considerations inevitably suggest that there must be an optimum mutation rate for the survival of a species under a given rate of environmental change. If the mutation rate is too high the species will be crushed under a heavy mutational load; if it is too low the species will not be able to cope with adverse environmental changes. The species that have managed to survive up to the present must be such that have been able to adjust their mutation rate to the optimum level through inter-group as well as intra-group selection.'

Kimura derived formulas to predict this optimum mutation rate, and then revisited the subject in a second paper, in which he considered the optimum mutation rate in a slowly changing environment (Kimura, 1967). At certain point he realized however that his formulas were too deterministic, because they did not incorporate the effect of drift. When he plugged the effect of drift into the equation, he was in a for a surprise. It turned out that 'random sampling of gametes has a very significant effect on the substitutional load' (Kimura and Maruyama, 1969).

The neutral theory and genetic variation <u>within</u> populations

In his 1968 paper Kimura noted in passing that genetic drift could not only account for the genetic differences *between* species, but also for the genetic differences *within* species: 'The fact that neutral or nearly neutral mutations are occurring at a rather high rate is compatible with the high frequency of heterozygous loci that has

been observed recently by studying protein polymorphism in human and *Drosophila* populations.'

In 1971 Kimura published, together with Tomoko Ohta, another paper on his budding neutral theory (Kimura and Ohta, 1971). In contrast to the 1968 paper, this paper focused on genetic variation within rather than between populations. It was a reply to critiques on his loose statement that not only genetic divergence but also polymorphism could be explained by drift.

In his 1968 paper, Kimura referred to the polymorphism studies of Lewontin and Hubby (1966) and Harris (1966). Lewontin and Hubby (1966) had studied 18 enzyme proteins in populations of *Drosophila pseudoobscura*, and found that 30 percent of the loci were polymorphic, with a heterozygosity of 12 percent. Harris (1966) studied human populations and found remarkably similar estimates: 30 percent polymorphism, and 9.9 percent heterozygosity. The claim of Kimura (1968) that these findings suggested that most mutations were neutral, had attracted two main objections.

One objection was that the data indicated that isolated populations often contained the same alleles. This observation was in agreement with balancing selection but not with drift, especially considering that apart from sharing the same alleles, isolated population also contained those alleles in similar frequencies. The second objection was that the observed genetic variation (measured as either heterozygosity or number of alleles per site) in large populations seemed to be lower than expected based on neutral dynamics, but were again in agreement with expectations based on balancing selection (Kimura and Ohta, 1971).

In reply to the first objection, Kimura and Ohta argued that gene flow is ubiquitous in highly mobile species such as *Drosophila*, mice and humans, leading to panmixia and hence to similarities in gene pools across populations. Kimura and Ohta investigated the validity of the second objection by mathematically deriving the expected heterozygosity given effective population sizes and assuming neutral forces (Kimura and Ohta, 1971). They did so by walking the opposite way: by estimating historic effective population sizes (Ne) of mice and humans from heterozygosity estimates. Based on available data on protein polymorphism across species (compiled by Selander et al., 1970), Kimura and Ohta assumed an average heterozygosity of 0.1. Next, they referred to a formula which Kimura published in a

previous paper ($He = 4 \cdot Nu_g/(4 \cdot Ne \cdot u_g + 1)$) (Kimura and Crow, 1964) to conclude $4 \cdot Ne \cdot u_g \approx 0.1$, and hence $Ne \cdot u_g \approx 0.025$. All that was left to do in order to derive Ne was to divide 0.025 by u_g.

Kimura and Ohta obtained the mutation rate per site per generation (u_g) from available data on mutation rate per site per year (u_s). They did so by correcting for generation time, the number of years per generation. In their own words: 'For species such as the mouse, with possibly two generations per year, the mutation rate per generation [...] is half as large [as the mutation rate per year], while for man it should be some twenty times as large.' (Kimura and Ohta, 1971) By plugging the obtained estimates of u_g into the equation $4 \cdot Ne \cdot u_g \approx 0.1$, they arrived at the conclusion that the (historical) effective population sizes (Ne) equalled 500,000 for mice and 13,000 for humans. As these estimates seemed to be roughly in accordance with reality, Kimura and Ohta argued that the observed genetic variation in natural populations was not in conflict with neutral theory expectations.

The mainstream perception at the time held that the divergence of populations and the polymorphism within populations were driven by two different types of selective processes. Divergence was thought to result from positive selection, whereas polymorphism arose through balancing selection. In contrast, Kimura and Ohta stated that divergence and polymorphism reflected two sides of the same coin: 'In our view, protein polymorphism and molecular evolution are not two separate phenomena, but merely two aspects of single phenomenon caused by random frequency drift of neutral mutants in finite populations.' (Kimura and Ohta, 1971)

If both polymorphism and divergence were indeed two aspects of a single phenomenon (being random drift of neutral mutations), a strong correlation between both aspects was to be expected because under neutrality both the level of polymorphism ($\theta = 4 \cdot Ne \cdot u_g$) and the level of divergence ($k = u_g$, k = substitution rate) depend on the mutation rate. A positive correlation between polymorphism and divergence was indeed observed, providing additional support for the neutral theory.

Two sides of the same coin?

The statement that genetic polymorphism largely reflects genetic drift of neutral mutations, has been criticized on several grounds. An early objection was based on the variation of levels of genetic polymorphism across loci within species (Lewontin and Krakauer, 1973). (In contrast, Lewontin's paradox, which I will discuss in the next section, is about variation of genetic polymorphism across species).

As mentioned above, as an argument against the neutral theory it was noted that isolated populations seem to contain the same alleles in similar frequencies. Lewontin and Krakauer quantified the allele frequency differences of various genes across human populations by calculating Fst-values (Lewontin and Krakauer, 1973). They argued that if the differences in allele frequencies between populations were caused by demography and not by selection, all loci should have generally similar Fst-values. In contrast, they found significant heterogeneity in locus specific Fst-values. Lewontin and Krakauer argued that this heterogeneity demonstrated that at least some loci were affected by selection (Lewontin and Krakauer, 1973).

Lewontin and Krakauer's test (the LK test) was severely criticized and quickly fell out of use (Beaumont, 2005). But of lasting importance for future selection analyses was Lewontin and Krakauer's proposition that demography affects the entire genome whereas selection affects specific genomic regions only. (In their own words: 'While natural selection will operate differently for each locus and each allele at a locus, the effect of breeding structure is uniform over all loci and all alleles.'(Lewontin and Krakauer, 1973)) This assumption, the 'Lewonton-Krakauer axiom' (Hahn, 2008), has become the implicit assumption of present day genome wide selection scans, which search for genomic regions which stand out from genome wide averages.

Another objection against the neutral theory revolves around the observed positive correlation between levels of polymorphism and levels of divergence. Begun et al. (2007) published the 'first true population genomic dataset' (Hahn, 2008), a dataset containing entire genomes of multiple individuals belonging to the same species (*D. simulans*). This dataset indicated that, contrary to previous belief, a positive correlation between polymorphism and divergence does in fact not exist. Instead, a comparison of genetic polymorphism within *D. simulans* and genetic divergence between *D. simulans* and *D. melanogaster*, showed a negative rather than

a positive correlation: genomic regions with high between species divergence, contained less within species variation (Hahn, 2008). As the neutral theory predicts a positive correlation between polymorphism and divergence, the finding of Begun et al. (2007) is at odds with the neutral theory.

Lewontin's paradox

Implicit in their calculations, and as they explicitly noted at the end of their 1971 paper, Kimura and Ohta provided a potential explanation for another puzzling observation: the relatively constancy of levels of genetic variation (He) across species. Data comparison (Selander et al., 1970) not only showed that the average heterozygosity across species was close to 0.1, but also that the variation around the mean was low: all species had an average heterozygosity close to 0.1. This could be considered surprising given the wide variation of species traits, including the supposedly relevant traits such as generation time and effective population sizes (after all: $He = 4 \cdot Nu_g/(4 \cdot Ne \cdot u_g + 1)$).

Kimura and Ohta argued that the observed uniformity of levels of genetic variation resulted from the inverse relation between population size and generation time. Humans had a generation time of 20 years and a Ne of 13,000, whereas mice had a generation time of 0.5 years and a Ne of 500,000. The net outcome was that both species had the same level of heterozygosity. 'The species with short generation time [and hence lower u_g] tends to have small body size and attain a large population number, while the species which takes many years for one generation [and hence has higher u_g] tends to have a small population number.' Therefore, the product $4 \cdot Ne \cdot u_g$ 'should be less variable among different organisms than its components.' (Kimura and Ohta, 1971)

Many population geneticists, neutralists and selectionists alike, do not consider this explanation satisfactory, and the absence of correlation between population size and theta is still known as the 'paradox of variation' or 'Lewontin's paradox' (Corbett-Detig et al., 2015; Hahn, 2008; Lewontin, 1974).

The *nearly* neutral theory of evolution

On the last page of their 1971 paper, Kimura and Ohta addressed an apparent discrepancy between theory and facts. Their estimates of average rate of amino acid

substitutions per generation (u_g) ($0.5 \cdot 10^{-7}$ for mice and $2 \cdot 10^{-6}$ for humans) were considerably lower than 10^{-5}, which they quoted as the standard figure (without providing a reference). This led Kimura and Ohta to suggest an important modification to the neutral theory: '10^{-7} *per year* is much lower than the standard figure of 10^{-5} *per generation* [...] and this suggests that, in general, neutral mutants constitute a small fraction of all the mutants [...]. Thus, we consider this as one important revision to earlier work.' (Kimura and Ohta, 1971)

This revision would in subsequent years be elaborated upon by Ohta but not so much by Kimura, leading to two opposing theories: the neutral theory of molecular evolution (Kimura's), and the *nearly* neutral theory of molecular evolution (Ohta's). Kimura's neutral theory holds that the majority of mutations fall into two categories. Either mutations are strongly deleterious or they are neutral. Otha's *nearly* neutral theory, in contrast, assumes that the majority of mutations fall into three categories: strongly deleterious, mildly deleterious and neutral. Both theories agree that the neutral mutations are mostly responsible for the observed genetic variation within populations. Kimura and Ohta: 'We must emphasize, however, that most mutants that spread into the species are neutral, even if the neutral mutants constitute a small fraction of all the mutants.' (Kimura and Ohta, 1971)

A main difference between the neutral and the nearly neutral theory concerns considerations around the relationship between u_y and u_g (i.e. the mutation rate per site per year vs the mutation rate per site per generation). In their 1971 paper, Kimura and Ohta assumed, as discussed above, a linear and proportional relationship and calculated u_g using the formula $u_g = u_y \cdot g$, with g denoting generation time (measured in years per generation). The number of germline DNA replication events is however not proportional to generation time, and therefore the relationship between u_g and u_y is not necessarily proportional. u_g/g (= u_y) could be smaller for species with long generation times compared to species with short generation times. If so, the speed of the molecular clock would depend on the generation time. More precisely, if would run slower in species with longer generation times. This hypothetical effect is called the 'generation-time effect'.

A generation-time effect was not apparent from the earliest protein studies. According to the 'genetic equidistance rule' (discussed below) the extent of genetic divergence between any two species depends almost exclusively on the time to the most recent common ancestor (TMRCA), and hence not on species traits. However, a study was published in 1969 in which the authors did report the detection of a generation time effect. They had detected differences in genetic divergence rates between rodents and primates (Laird et al., 1969). They wrote: 'The initial rate of nucleotide sequence variation among rodents is ten-fold higher than that among artiodactyls when divergence time is estimated in years. This difference diminishes if generations, rather than years, represent the appropriate interval of evolutionary divergence.' (Laird et al., 1969) Interestingly, and perhaps tellingly, whereas previous studies analysed proteins, Laird et al. (1969) analysed non-coding DNA. This suggested that non-coding DNA exhibited a generation time effect, whereas coding DNA did not (Ohta, 1995).

Why should non-coding DNA behave differently from coding DNA? Ohta's nearly neutral theory provides a potential explanation. Ohta (1992) argued that the absence of the generation time effect in proteins was ultimately caused by the interplay between selection and drift. Her reasoning was based on the assumption that mutations in protein coding sites are not neutral but instead slightly deleterious. One consequence is that fixation probabilities for these sites do not equal the inverse of the effective population size, and therefore that substitution rates for these sites do not equal mutation rates (i.e. $k \neq u_g$). Instead, the fixation probability of these slightly deleterious mutations are determined by the selection coefficient and the effective population size. The mutations are effectively eliminated in large populations, but less effectively in small populations, in which drift is more dominant.

Species with short generation times tend to have big population sizes. In contrast, species with long generation times generally have relatively low effective population sizes and thus experience more genetic drift. As a result, a certain proportion of slightly deleterious alleles fixates despite being negative selected against. The net effect is that although over time species with short generation times accumulate more mutations in protein coding sites than species with long

generation times, substitution rates will vary to a lesser extent. (Empirical evidence for this hypothetical mechanism has been reported by Galtier, 2016).

More recent and more refined investigations have analysed variation in substitution rates among synonymous and non-synonymous sites. If Ohta's (1992) reasoning is correct, a generation time effect should be present in synonymous sites but less so in non-synonymous sites, and this is indeed what has been reported by several studies (Ohta, 1993; Wu and Li, 1985). However, other studies have found contrasting results, pointing both at the presence of a generation time effect in non-synonymous sites (Thomas et al., 2010), and conversely on the absence of a generation time effect in synonymous sites (Kumar and Subramanian, 2002).

The molecular clock

As discussed above, the original motivation for Kimura to develop his neutral theory of molecular evolution was Haldane's dilemma. In his view, the neutral theory solved the mismatch between expected substitution rates and observed substitution rates by stating that most substitutions were driven by drift rather by than selection. By doing so, Kimura demonstrated strong confidence in Haldane's reasoning. The alternative and perhaps more obvious implication of the mismatch between observations and expectations was that Haldane's theoretical framework was flawed.

Many authors have indeed argued that Haldane's upper limit of 1 substitution per 300 generations is incorrect (Brues, 1964; Dodson, 1962; Felsenstein, 1971; Maynard-Smith, 1968; Sved, 1968; Van Valen, 1963). These critiques came out both before and after Kimura published his 1968 paper and were based on various grounds. One famous objection is that Haldane's upper limit of 1 substitution per 300 generations seems to be too restrictive when compared to the numerous phenotypic differences between species (Dodson, 1962). Haldane's limit allows for example for around 800 selective driven substitutions in the human lineage since the split from chimpanzees, a number which seems low compared to the observed phenotypic differences between both species. More recently, Haldane's limit has also been challenged by simulation studies which provide evidence for substitutions rates several orders of magnitudes higher than Haldane's limit of 1 substitution per

300 generations (Hickey and Golding, 2019; Nunney, 2003; Weissman and Barton, 2012).

But although Kimura's original argument was based on the concept of the cost of selection, the neutral theory has become independent of cost of selection considerations, such that a refutation of Haldane's limit does not equal refutation of the neutral theory (Ohta and Gillespie, 1996). Even in 1968, just months after Kimura published his paper and before Kimura provided additional evidence for the neutral theory in later papers, Maynard-Smith wrote in reply: 'I do not want to query the conclusion that drift has been important, but Kimura's conclusion that the rate of evolution is too great to be explained by natural selection can be queried.' (Maynard-Smith, 1968).

The independency of the neutral theory of Haldane's cost of selection argument arises from the fact that the theory rests on multiple lines of evidence. Arguably the most important of them, at least when only considering arguments from within the field of population genetics, is the molecular clock hypothesis. Even though Kimura developed his theory on different grounds, he soon regarded the existence of a molecular clock as the main fundament of the neutral theory. Kimura and Ohta (1971) wrote in the beginning of their paper: 'Probably the strongest evidence for the theory is the remarkable uniformity for each protein molecule in the rate of mutant substitutions in the course of evolution.'

The existence of a molecular clock was first discovered through analyses of single proteins. Margoliash (1963), studying cytochrome c, concluded: 'It appears that the number of residue differences between cytochrome c of any two species is mostly conditioned by the time elapsed since the lines of evolution leading to these two species originally diverged.' From this observation Margoliash (1963) induced the 'genetic equidistance rule', stating that all species pairs with a similar TMCRA have a similar genetic distance. Zuckerkandl and Pauling (1965) independently arrived at the same conclusion by studying a different gene, namely haemoglobin. They found that although a pair of species differed on average more in their heamoglobin sequence than in their cytochrome c sequence, the genetic equidistance rule held true for heamoglobin as well: the number of heamoglobin sequence difference for any species pair was proportional to the TMRCA of the species pair. Zuckerkandl and Pauling (1965) argued that the genetic equidistance

rule implies a constant rate of molecular change (i.e. substitution rate) over time, a conjecture now known as the molecular clock hypothesis (Wilson and Sarich, 1969; Zuckerkandl and Pauling, 1965).

The differences in substitution rate across proteins has been attributed to differences in protein functional constraints, resulting in different proportions of deleterious and neutral (i.e. silent and conservative) mutations (Dickerson, 1971). The constancy of substitution rates within proteins across lineages provided evidence for the (near) absence of advantageous mutations, because there is no reason to believe that adaptive changes would occur in a clock-like manner. Mutation and random drift, on the other hand, make for a plausible mechanism behind constant substitution rates. Selectionists argue however that the molecular clock hypothesis has been proven incorrect, and that the data shows that substitution rates vary considerably, both over time and between lineages (see (Kern and Hahn, 2018).

The existence of a molecular clock was first induced from and confirmed using data on just a handful of proteins (e.g. Dickerson, 1971). When in subsequent years data on protein and DNA sequences accumulated, it became increasingly clear that the perception of a steady and constant clock was an oversimplification. In modern day phylogenetics, a fixed clock is known to causes bias in estimates of divergence times as well as bias in inferences of deep relations (Drummond et al., 2006). The standard practice when constructing phylogenetic trees is therefore to apply relaxed clocks rather than a fixed clock. These relaxed clocks allow for rate heterogeneity, both across lineages and across clades.

But does the existence of rate heterogeneity reject the neutrality model? If substitution rates are driven by stochastic processes, some stochastic variability in substitutions rates among lineages and through time is to be expected. When are observed deviations too strong to fit neutral model expectations? Theoretically, this question can be answered by generating a probability distribution of number of substitutions per time interval (i.e. branch length in years) given an average substitution rate.

An early attempt to statistically test the constancy of the molecular clock was published in 1974 (Langley and Fitch, 1974). The underlying assumption of the study was that the expected probability distribution of substitution rates, given

neutral dynamics, resembles a Poisson distribution. The authors, Langley and Fitch, compiled the available data on amino acid sequences of four proteins (alpha and beta hemoglobins, cytochrome c, fibrinopeptide a) for vertebrate species, and calculated the minimum number of nucleotide substitutions required to explained the observed amino acid differences (using the 'minimum phyletic distance method', Langley and Fitch, 1974) given an a priori phylogenetic vertebrate tree. Next they compared these observed estimate to expected values, calculated with a Poisson function (i.e. number of occurrences given number of years, given the average mutation rate). They tested the goodness of fit using a chi-squared test and a likelihood ratio test. Both methods returned highly significant p-values, leading them to reject the null model of constant overall substitution rates, and to conclude that substitutions rates vary both within genes across lineages and within lineages across genes.

Some aspects of the Langley and Fitch (1974) study were however criticised by Hudson (1981). Hudson argued that the analysis of Langley and Fitch (1974) did not truly test the neutral theory, which he referred to as the 'constant-rate neutral model' (Hudson, 1981). Hudson argued that, as acknowledged by Langley and Fitch (1974), their chi-squared distribution was positively correlated to the population genetic parameter θ, defined as $4 \cdot Ne \cdot u$. Therefore, as Hudson pointed out, 'no matter how large the observed value of X^2_{LF}, a sufficiently large value of θ could account for the observation.' In here X^2_{LF} denoted the chi-squared test values obtained by Langley and Fitch (1974).

Hudson therefore performed Monte Carlo simulations to determine the exact relationship between X^2_{LF} and θ, which in turn he used to estimate the values of θ needed to explain the observed X^2_{LF}. Hudson found that the observed X^2_{LF} could only be explained with θ values higher than 10, which, as Hudson pointed out, were 'incompatible with the low levels of heterozygosity observed at the hemoglobin loci in humans'. Hudson arrived therefore at the same conclusion as Langley and Fitch (1974): 'The constant-rate neutral model is highly improbable. Other neutral models and models involving natural selection need to be considered.'

The disparity between expected rates and observed rates can be quantified using the dispersion index, which is the ratio (R)of the variance (V)of the number of substitutions on a lineage to the mean (M) number per lineage. One of the

characteristics of a Poisson distribution is that the variance V is equal to the mean M, and therefore that the ratio equals one (i.e: R = V/M = 1). Gillespie (1989) analyzed 20 proteins and found R values ranging between 0.32 and 43.82, of which 12 values were significantly above 1, as was the average. Several hypotheses have been proposed since to account for an overdispersed molecular clock (Ayala, 2000; Cutler, 2000; Bedford and Hartl, 2008). One explanation is that overdispersion results from purifying selection (Cutler, 2000), in which case overdispersion of the molecular clock would not be at odds with the neutral theory.

In the wake of the disparity index, several other methods have been proposed to test the existence of a global molecular clock. From these tests it became increasingly clear that the perception of a steady and constant clock was an oversimplification. This finding called for new methods were developed to estimate divergence times in the absence of a molecular clock.

Divergence times are easy to calculate when assuming a global molecular clock with known fixed rate. Divergence times are difficult to calculate if assuming that the substitution rate can differ among lineages. To do so one first needs to convert an additive tree into ultrametric (linearized) tree. In other words: convert a tree with branch lengths proportional to the number of substitutions, to a tree in which nodes are dated, and in which all branches line up. Langley and Fitch (1974) had presented an early method. In the 90's this method was surpassed by the non-parametric rate smoothing approach (NPRS, Sanderson, 1996), with other methods such as penalized likelihood (PL) and Bayesian methods (Thorne et al. 1998) following shortly (Britton et al., 2007). The use of fixed global clocks fell out of use, and was replaced by 'relaxed phylogenetics', in which phylogenies are inferred using either local clock models of models in which rates vary across lineages in an autocorrelated manner (Yoder and Yang, 2000; Drummond et al., 2006). (Note: this is different from rate heterogeneity across sites, which is controlled by selecting the appropriate substitution model.)

In his monograph on the subject, Kimura stated that 'emphasizing local fluctuations as evidence against the neutral theory, while neglecting to inquire why the overall rate is intrinsically so regular or constant is picayunish. It is a classic case of 'not seeing the forest for the trees' (Kimura, 1983). Even so, it is an interesting

and legitimate question to ask to what extent deviations from a steady molecular clock reflect stochastic variability or instead perturbations due to selective events.

As I will discuss in the next section, proteins consist of both functional and non-functional sites. Deviations from constant rates are expected for functional sites but not for non-functional (and hence neutral) sites. The constancy of the molecular clock has indeed been confirmed when studying neutral sites only. Kumar and Subramanian (2002) analysed fourfold degenerative sites in a mammalian dataset of over 5000 genes and found that divergence time strongly correlates with evolutionary distance (number of nucleotide differences), with a correlation coefficient of 0.97 (but for a contrasting result, see: Green et al., 2014).

Evidence for the neutral theory from molecular biology

The main conjecture of the original neutral theory is that most mutations are either neutral or strongly deleterious (Kimura, 1968; Kimura and Ohta, 1971). Kimura arrived at this conclusion using considerations from the field of theoretical population genetics. Arguably, the most intuitive evidence in favour or against neutral molecular evolution is however not to be found within the field of population genetics, but within the field of molecular biology.

The first scientists to systematically pursue this line of evidence were Thomas King and Thomas Hughes Jukes, who published a paper on the subject in 1969 (King and Jukes, 1969). In here, King and Jukes provide an in depth analysis of DNA and proteins to arrive at the same conclusion as Kimura, namely that 'the stream of spontaneous alternations in DNA, constantly fed into the genetic pool, should include far more acceptable changes that are neutral than changes that are adaptive' (King and Jukes, 1969).

King and Jukes' 1969 paper was built around several arguments. One argument was that due to the redundancy of the genetic code, many substitutions within genes do not result in an amino acid replacement and therefore are silent or synonymous. Of all possible single nucleotide substitutions, 25 percent are non-synonymous, and yet they represent the vast majority of actual substitutions within genes. King and Jukes (1969) argued that this observation implied that most non-synonymous mutations are deleterious, as stated by the neutral theory.

A second argument was that even non-synonymous substitutions might not cause differences in the structure and/or functionality of proteins. These so called conservation substitutions, of which King and Jukes (1969) provided numerous examples, are therefore effectively neutral, just like synonymous mutations.

A third argument involved the relative frequencies of each type of amino acid within proteins (i.e. amino acid composition). King and Jukes (1969) showed that these relative frequencies were proportional to the number of redundant DNA codons coding for each amino acid. For example, the amino acid serine, which is coded for by six DNA codons (TCT, TCC, TCA, TCG, AGT, AGC) is three times more abundant than the amino acid lysine, which is coded for by two DNA codons (AAA, AAG). This correlation is expected if neutral forces are at play, but hard to explain from a selectionist's point of view.

The argumentation of King and can thus be summarized as follows. Substitutions in protein coding regions are mostly either deleterious (non-conservation non-synonymous substitutions) or neutral (synonymous substitutions or conservation substitutions). The low proportion of positively selected sites within protein-coding regions is evident from the disproportionally high ratio of synonymous vs non-synonymous substitutions as well as from the amino acid composition of proteins.

King and Jukes (1969) on non-coding DNA

King and Jukes (1969) also mentioned a potential shortfall in Kimura's reasoning which so far had not been spotted by other authors. Kimura (1968) had extrapolated the observed substitution rate in a few proteins to derive an estimate of the substitution rate genome wide. Because 'Kimura's argument was deliberately conservative in some respects', King and Jukes (1969) redid some of the calculations and arrived at an estimate of 'about two allele substitutions per year', slightly higher than Kimura's estimate of one substitution per two years. (In contrast to Kimura (1968), King and Jukes (1969) refer to amino acid substitutions rather than nucleotide substitutions. Because one amino acid substitution corresponds to ~1.25 nucleotide substitutions, these estimates are roughly similar.)

King and Jukes (1969) then argued that both estimates appeared 'much too high', because these substitution rates would be associated with high mutation loads

(which is different from Kimura's 'genetic load'). Reconsidering their equations, King and Jukes (1969) realized that they had overlooked something – and Kimura (1968) as well. The calculation was based on 'the assumption that all or most mammalian DNA consists of structural genes'. But, as King and Jukes (1969) pointed out, both theoretical considerations and empirical evidence indicated that most genomes contained less than 40.000 genes. Given that proteins consisted generally of only a few hundred or thousands amino acids, it appeared that only a small proportion of the genome – 'not much more than 1 percent' – coded for proteins. When multiplying the estimated substitution rate per codon with the combined gene length rather than with the genome length, King and Jukes (1969) arrived at a much lower estimate of the substitution rate: 'If the average gene consists of 1000 nucleotide pairs, extrapolation from the estimated $16 \cdot 10^{-10}$ substitutions per codon per year gives one amino acid substitution per species per 50 years. This is a far more believable figure.'

The conclusion that the majority of the genome did not code for proteins, was a side finding, but a remarkable and relevant finding nevertheless. King and Jukes (1969) wrote: 'Either 99 percent of mammalian DNA is not true genetic material, in the sense that it is not capable of transmitting mutational changes which affect the phenotype, or 40.000 genes is a gross underestimate of the total genome.'

The idea that the majority of the genome does not code for proteins is now part of mainstream thought (Lander et al., 2001). Unbeknownst to King and Jukes, a fraction of non-protein coding DNA operates in regulating gene expression, but this does not alter the overall conclusion that large proportions of the genome are non-functional and therefore neutral. This conclusion is nowadays occasionally used as an argument in favour of the neutral theory (Jensen et al. 2019). However, this was not how it was intended by King and Jukes (1969), who strived to obtain a better estimate of amino acid substitution rates.

The neutral theory was developed in a time when actual sequence data was limited to proteins, and meant to explain observed patterns in these data sets. As a consequence, the original arguments for the neutral theory pertained to proteins, not to full genomes. This is reflected in the abstract of King and Jukes (1969), which reads: 'Most evolutionary change in proteins may be due to neutral mutations and genetic drift.' From a historical perspective, it is therefore incorrect to argue that the

main proposition of the neutral theory – that most substitutions are nonadaptive – is true simply because the majority of the genome is non-coding.

Genetic draft

Selectionists do not question the neutrality of intergenic sites, only the neutrality of mutations that affect the phenotype (Hahn, 2008). Still, they do argue that even non-functional sites might not be free from the influence of selection, as selection on functional sites might indirectly affect the genetic variation of non-functional neighbouring sites, namely through linkage.

The process of indirect selection driven allele frequency change through linkage is known as genetic draft (Gillespie, 2000). When genetic draft causes a frequency increase of neutral or slightly deleterious alleles due to linkage to beneficial alleles, it is also known as genetic hitchhiking (Maynard-Smith and Haigh, 1974). The opposite, the removal of neutral or slightly beneficial mutation linked to deleterious alleles, is known as background selection (Charlesworth, 2012; Charlesworth et al., 1993). Genetic draft, whether through a selective sweep or through background selection, has the potential to influence both genetic divergence and genetic polymorphism, and as such cause deviations from values predicted by the neutral theory.

The extent to which genetic draft causes deviation in divergence levels depends on the relative proportions of neutral and non-neutral mutations which are affected by genetic draft. Because genetic draft has the same net effect as genetic drift – i.e. the random fixation or loss of a mutation – linkage to advantageous or deleterious mutations affects the substitution rate of non-neutral mutations (i.e. decreased substitution rates of advantageous mutations and increased substitution rates of deleterious mutations), but not the substitution rates of neutral mutations (Birky and Walsh, 1988). In non-coding regions, where mutations are neutral, linkage should therefore not cause deviations from substitution rates predicted by the neutral theory (Jensen et al., 2019).

The extent to which genetic draft causes deviation in divergence levels largely depends on recombination rates (Cutter and Payseur, 2013): the higher the recombination rates, the narrower the window of selective sweeps and background selection. As both a selective sweep and background selection lead to reduced

variation in genomic regions surrounding a positively or negatively selected allele, a relationship is to be expected between recombination rates and polymorphism levels. Multiple studies have indeed reported that genomic regions with low recombination rates have less genetic variation than genomic regions with high recombination rates (Begun and Aquadro, 1992; Corbett-Detig et al., 2015; Cutter and Payseur, 2013; Lohmueller et al., 2011). Another observed trend is that the genetic variation within genomes is lowest close to coding or conserved non-coding sites (Cutter and Payseur, 2013; Lohmueller et al., 2011).

Selectionists argue that the correlation between recombination rates and genetic variation may provide an explanation for Lewontin's paradox, the observed narrow range of levels of genetic diversity across species with widely different population size (Corbett-Detig et al., 2015; Hahn, 2008; Kern and Hahn, 2018). They also argue that the correlation disagrees with the neutral theory (Hahn, 2008; Kern and Hahn, 2018). Neutralists however point out that the observed correlation might be mostly due to purifying selection (background selection) rather than to positive selection (selective sweep, Lohmueller et al., 2011), and therefore are in fact consistent with the statement that most mutations are either neutral or deleterious (Jensen et al., 2019).

The mathematical work of Kimura and Ohta (Kimura and Ohta, 1971) on expected levels of polymorphism (captured in the formula: theta = $4 \cdot Ne \cdot u_g$), did not consider the effect of genetic draft, as it assumes sites to be unlinked. The alternative to a rejection of the neutral theory, if mainly based on this ground, would be to incorporate the effect of genetic draft into the neutral theory, thereby creating a model which accounts for the indirect effects of purifying selection (Comeron, 2017; Jensen et al., 2019).

McDonald-Kreitman test

The molecular clock hypothesis and the neutral theory were both proposed in a time that actual DNA-sequence data was non-existent. The molecular clock hypothesis was based on amino acid sequence data, not on nucleotide sequence data. So was the neutral theory. Kimura (1968) derived his estimate of 1.8 nucleotide substitutions per year indirectly from data on amino acid substitutions. He reasoned: 'Because roughly 20 per cent of nucleotide replacement caused by

mutation is estimated to be synonymous, [...] one amino-acid replacement may correspond to about 1.2 base pair replacements in the genome.'

One of the first studies which analysed genetic variation within a natural population using actual DNA-sequences rather than using amino acid sequence data was published in 1983 by Martin Kreitman (Kreitman, 1983). Kreitman (1983) sequenced the alcohol dehydrogenase (Adh) gene of 11 *Drosophila melanogaster* individuals, belonging to 5 different populations divided over continents. Within coding regions, Kreitman found 14 polymorphic nucleotide sites, of which 1 resulted in a polymorphic amino acid. So although only 25 percent of all possible nucleotide substitutions are synonymous, they represented (13/14 =) 93 percent of observed polymorphic sites. Kreitman: 'The implication is that most amino acid changes in Ahd would be selectively deleterious.' (Kreitman, 1983)

The nature of the non-synonymous substitution was however unclear. Any non-synonymous substitution can either be a conserving substitution, a slightly deleterious substitution which was overruled by drift, or a positively selected substitution. In 1987 and 1991 Kreitman published another two papers on Ahd gene sequence analysis (Hudson et al., 1987; McDonald and Kreitman, 1991). In both papers the authors introduced a method to test whether observed non-synonymous substitutions were driven by selection.

The HKA (Hudson, Kreitman, Aguade) test compared levels of polymorphism (number of segregating sites) and divergence (number of substitutions) among two genomic regions, such as the Adh coding region and a flanking region. The rationale behind the test is that different trends across loci can be indicative of selection.

Although the HKA test is still being used (e.g. Liu et al., 2014), the second test introduced by Kreitman, the McDonald-Kreitman (MK) test (McDonald and Kreitman, 1991), is the more popular among the two. The MK test involves an analysis of a single gene, rather than a comparison between multiple genes as for the HKA test. The test is based on a comparison between the number of segregating and fixed degenerate and non-degenerate sites.

McDonald and Kreitman (1991) argued that if mutations in non-degenerate sites (i.e. non-synonymous mutations (N)) are regulated by neutral dynamics only, they should have equal probabilities of fixation or loss as mutations in degenerate sites (i.e. as synonymous mutations (S)), as well as equal times of segregating before

going to fixation or loss. Therefore, we would expect the ratio N_{fix}/S_{fix} (= $\omega_{divergence}$) to equal the ratio N_{poly}/S_{poly} (= $\omega_{polymorphism}$). As mutations under positive selection fixate quickly, positive selection on N mutations would result in: $N_{fix}/S_{fix} > N_{poly}/S_{poly}$. Puryfing selection on N mutations would result in: $N_{fix}/S_{fix} < N_{poly}/S_{poly}$.

McDonald and Kreitman (1991) compared the Adh locus of three *Drosophila* species, and found overall 44 segregating sites, containing 42 synonymous and 2 non-synonymous mutations. They also found 24 fixed differences between the species, of which 17 were synonymous and 7 non-synonyymous. In other words, around 29 percent of the fixed differences between species were non-synonymous, whereas this category made up only 5 percent of the total number of segregating sites. The deviation from equal ratios was significant (according to a G-test of independence), providing compelling evidence for positive selection.

Because relaxation of purifying selection can cause the MK test to wrongly infer positive selection, it is advised to first compare the polymorphism levels between species (He et al., 2018).

dN/dS tests

Both the HKA test and the MK test require genetic data not only from two species, but also for multiple individuals for at least one of those species. Many genetic datasets don't meet these requirements, because they contain either data on one individual for multiple species, or data for multiple individuals for one species. To be able to test for neutrality in these types of datasets, other methods have been developed.

Fumio Tajima published a paper in 1989 in which he introduced a method to test for neutrality based solely on patterns of genetic variation within populations/species (Tajima, 1989). This method tests for deviation from the expected relationship between the number of segregating sites and nucleotide diversity (i.e. the average number of differences between two randomly drawn sequences) (Watterson, 1975). Modifications have been published in subsequent years (Fay and Wu, 2000; Fu and Li, 1993), but Tajima's D has remained popular, as it is a relatively robust and easy test to apply.

Whereas Tajima's D is based solely on polymorphism data, another test originally developed in the 1980's, the dN/dS test, is based solely on divergence

data. Another difference is that Tajima's D test can be applied to both protein coding and non-protein coding data, whereas the dN/dS test requires protein coding data.

dN is the ratio between the actual number of non-synonymous substitutions and the potential number of non-synonyymous substitutions. Similarly, dS is the ratio between the actual number of synonymous substitutions and the potential number of synonymous substitutions. Several methods have been proposed to obtain reliable estimates (Li et al., 1985; Nei and Gojobori, 1986).

It was initially reasoned that a dN/dS ratio of 1 implies the gene is evolving neutrally, whereas values below 1 indicate purifying selection and values above 1 positive selection (Hughes and Nei, 1988). However, most non-synonymous mutations are under puryfing selection (Hughes et al., 2003), and therefore observed dN/dS ratios are generally below 1, with typical gene wide average values of dN/dS ranging between 0.2 and 0.3, meaning that on average 80-90 percent of non-synonymous mutations are deleterious (Fay et al., 2001; Mugal et al., 2014; Nielsen and Yang, 1998). As a consequence, positively selected nonsynonymous mutations, which comprise the minority of nonsynonymous mutations, will be overlooked when assessing gene wide averages.

The identification of this problem has led to the development of more sophisticated dN/dS tests with increased power. These tests execute sliding window analyses of dN/dS, rather than calculating a single estimate for the entire gene (Nielsen and Yang, 1998). The obtained distribution of dN/dS values across the gene is used to test the performance of so called 'site-models'. The null model holds that the obtained dN/dS ratios (also denoted omega ω) across the gene are either 1 or negative (reflecting respectively drift and purifying selection). The alternative model holds that in addition at least one region within the gene has a dN/dS ratio above 1, indicative of positive selection.

Whereas dN/dS tests and MK tests were initially executed on single genes (Hughes and Nei, 1988; McDonald and Kreitman, 1991), now it is routine practice to analyse datasets containing thousands of genes. Although some studies still apply candidate gene approaches (Liu et al., 2010; Xu et al., 2013), these whole genome comparisons have become mainstream (Agaba et al., 2016; Foote et al., 2015; Kumar et al., 2015; Tsagkogeorga et al., 2015; Zepeda Mendoza et al., 2018).

The rate of adaptive molecular evolution (alpha)

Apart from providing an objective way to test for deviations from neutrality, the MK test and the dN/dS tests also provide ways to estimate the proportion of neutral, advantageous and deleterious non-synonymous mutations.

When assuming that there are no advantageous non-synonymous mutations, the proportion of neutral non-synonymous mutations simply equals the dN/dS ratio (ω), and the proportion of deleterious non-synonymous mutations simply equals 1 – ω. For example, a typical dN/dS value of 0.2 indicates that 20 percent of non-synonymous mutations behaved similar to synonymous mutations, whereas 80 percent of non-synonymous mutations have been under purifying selection (Eyre-Walker and Keightley, 2007).

The proportion of adaptive substitutions is known as alpha (α), or the rate of adaptive molecular evolution (Smith and Eyre-Walker, 2002), and can be derived, using a method based on the MK test, as follows: $\alpha = 1 - (S_{fix}*N_{poly})/(N_{fix}*S_{poly})$. For the dataset analysed by McDonald and Kreitman, alpha equals: 1 - (17*2)/(7*47) = 0.88. This suggests that 88 percent of all non-synonymous substitutions within the Adh gene of the sampled *Drosophila* species were driven by selection.

The MK test traditionally divides synonymous and non-synonymous sites into segregating and fixed sites. MK test-based methods have been developed which don't consider just the frequency of fixed and segregating sites, but also the the minor allele frequencies (MAF) within segregating sites. This improvement was initiated by Fay et al (2001), who noted that the dN/dS ratio of sites with low minor allele frequencies (MAF < 0.05) is considerably higher than the dN/dS ratio of sites with higher minor allele frequencies. It is thought that this excess of non-synonymous sites with low MAF is caused by slightly deleterious mutations, which can segregate for a while at low frequencies before getting lost. As these mutations cause an increase in S_{poly}/N_{poly} without affecting S_{fix}/N_{fix}, they can lead to an underestimation of alpha. Some authors therefore omit from their calculations polymorphisms segregating at low levels. (Charlesworth and Eyre-Walker, 2008; Eyre-Walker and Keightley, 2009; Fay et al., 2001; Galtier, 2016).

Another important insight is that the application of the MK and dN/dS tests is not restricted to protein coding regions. Both tests compare patterns of functional and non-functional sites. The subdivision between functional and non-functionally

sites typically consists of a subdivision in non-synonymous (N) and synonymous sites (S) within a gene. But in principle any subdivision between coding and adjacent non-coding sites, such as conserved regions and non-conserved flanking regions, can be used as input to the MK or dN/dS test (Jenkins et al., 1995). Analyses on these types of datasets provide insight into the question of to what extent adaptive molecular evolution involves changes in protein coding regions and to what extent changes in regulatory sequences.

Distribution of fitness effects (DFE)

Traditionally, mutations have been divided into three categories: advantageous ($w>1$), neutral ($w=1$), and deleterious ($w<1$), where 'w' denotes fitness. If plotted as a frequency histogram, this would result in three bars, with, according to the neutral theory, two relatively high bars (for classes: $w<1$ and $w=1$), and one low bar (for class: $w>1$). Such a histogram of fitness effect is called a distribution of fitness effects (DFE). Eyre-Walker et al. (2006) introduced a new method to estimate this distribution. Whereas DFE had previously been estimated from mutagenesis and mutation accumulation experiments (Eyre-Walker and Keightley, 2007), the method proposed by Eyre-Walker and colleagues was based on polymorphism data. This method infers the DFE from deviation of the observed allele frequency distribution (i.e. site frequency spectrum (SFS)) of non-synonymous SNPs from the expected SFS based on neutral dynamics. The method produces a more fine-scaled DFE than the traditional 3-class categorization, and provides an alternative method to estimate the rate of adaptive molecular evolution (alpha). In contrast to the MK test, which requires both polymorphism and divergence data, the method of Eyre Walker et al (Eyre-Walker et al., 2006) is solely based on polymorphism data (Tataru et al., 2017).

Are most non-synonymous mutations indeed either neutral or deleterious?

The neutral theory states that 'most' mutations are either neutral or strongly deleterious, If taken literally, this means that alpha, the proportion of adaptive substitutions, is generally below 50 percent. Kimura has however also used the term 'overwhelming majority' instead of 'most' (Kimura, 1991), implying he envisioned

alpha to be closer to 0 percent. Which percentages have been estimated from genomic datasets, and how do these estimates reflect on the neutral theory?

Fay et al. (2001) analysed 182 orthologous humans and old world monkey genes and found estimates of $\omega_{divergence} = 0.34$ and $\omega_{polymorphism} = 0.2$, leading them to conclude that 35% of observed non-synonymous substitutions between humans and old world monkeys have been driven by positive selection. In 2005, following the completion of a whole genome sequence of a chimpanzee, Waterson et al (2005) repeated this analysis using a bigger gene set and with as reference species chimpanzee rather than old world monkeys. They obtained statistically indistinguisable values of $\omega_{divergence} = 0.23$ and $\omega_{polymorphism} = 0.21\text{-}0.23$, leading to an estimate of alpha close to 0 (Waterson et al., 2005).

What caused the difference between the outcomes of these two studies? Waterson and colleagues suspect that the estimate of Fay et al. (2001) was inflated due to methodological issues: 'Because the previous results involved comparison to Old World monkeys, it is possible that they reflect strong positive selection earlier in primate evolution; however, we suspect that they reflect the fact that relatively few genes were studied and that different genes were used to study polymorphism and divergence.' (Waterson et al., 2005) Further studies have indeed confirmed the near absence of adaptive substitutions in humans (Eyre-Walker and Keightley, 2009; Zhang and Li, 2005).

Meta-analyses have however indicated that alpha is not universally low across species, but instead varies widely. Estimated alpha values are close to zero in humans, other primates, giant Galapagos tortoise, yeast, fungi, and nine plant species, but above 0.5 in fruitfly, mouse, rabbit, sea squirt, sunflower and enterobacteria (Fay et al., 2001; Galtier, 2016; Gossmann et al., 2010). The observed variation in alpha values is thought to be related to differences in effective population size (Galtier, 2016; Gossmann et al., 2012).

Ignoring the variation between species, average values of alpha have been found to be above 0.5 or just under 0.5 for almost all studied groups of animals (i.e. mammals, reptiles, birds, arthropods, molluscs, echinoderms). These estimates of alpha do not correspond with Kimura's statement that the 'overwhelming majority' of mutations are either neutral of deleterious.

Concerns have been raised about the reliability of the estimates of alpha (Hughes, 2007; Nei et al., 2010). One potential flaw is in the assumption that sites are independent. Fay (2011) has argued that 'the common assumption of independence among sites must be relaxed before abandoning the neutral theory of molecular evolution'.

Neutralists have also argued that there is a certain circularity involved in concluding high proportions of adaptive mutations from studies focussing on functional regions. Functional sites – coding regions and regulatory sequences – make up a small and biased proportion of entire genome, and therefore provide a strongly inflated estimate of the proportion of adaptive driven substitutions across the genome (Jensen et al., 2019). n that sense, whole genome scans provided a more reliable way to estimate the obiquity of positive selection events within the genome.

From candidate gene studies to whole genome scans

During the 2000s, the reduction in sequencing costs and the associated increase of genetic data led to a shift from candidate gene studies to whole genome scans. Until then selection tests were executed using a top-down approach, in which the detection of phenotypic traits under selection triggers the search for the underlying genotypic variation. With the advent of next generation sequencing, the focus shifted from top-down to bottom-up approaches, in which the detection of genetic signals of selection triggers the search for associated phenotypic traits. This concept, also known as reverse ecology, directly uncovers the genetic basis of adaptation, rather than first identifying phenotypic traits that are acted upon by natural selection (Li et al., 2008; Lotterhos and Whitlock, 2015).

Whole genome scans can either be based on entire genomes or on reduced representation libraries (RRLs), which contain subsets of random or targetted subset of snps scattered throughout the genome. Initially mostly applied to human datasets (Akey, 2009; Akey et al., 2002), following the generation of new lab protocols such as RADseq (Baird et al., 2008), RRLs became also available for non-human and non-model species (review: Haasl and Payseur, 2016; example of an early paper: Hohenlohe et al., 2010; for a list of whole genome selection scan studies and their findings, see appendix A1).

Whole genome scans typically find that levels of genetic variation and levels of genetic divergence fluctuate stochastically across the genome. Plots with levels of genetic polymorphism/divergence on the y-axis and genomic position on the x-axis often resemble irregular seismic waves diagrams, even after applying smoothing methods. The assumption is generally that this noise reflects the backdrop of neutral variation from which selected loci might or might not stand out. The smaller the effective population size and the consequent larger effect of drift, the wider the backdrop of neutral variation, and the more difficult it becomes to discrimate loci under selection (Bamshad and Wooding, 2003).

Genome wide selection scans search for genomic regions under selection generally in either of two ways (for a more detailed classification, see (Oleksyk et al., 2010; Weigand and Leese, 2018):

- by searching within populations/species for genomic regions with depleted genetic variation; or
- by searching for genomic regions with accelerated levels of genetic differentiation between populations or species.

Both approaches can potentially be used in combination with the comparative method, which entails searching for consistent signals of selection in populations which have likely been under similar selection pressures (e.g. Hohenlohe et al., 2010; Parker et al., 2013).

Genomic regions with depleted genetic variation can be detected by generating locus specific or sliding window estimates of polymorphism measures such as theta (Diller et al., 2002), heterozygosity (Oleksyk et al., 2008), runs of homozygosity (ROH) length, Tajima's D, LD extent (Hawks et al., 2007) and haplotype length/frequency statistics.

The estimate used to detect genomic regions with elevated levels of genetic differentiation depends on the TMRCA (Sabeti et al., 2006). When comparing species (i.e. when searching for selection on derived mutations), genomic regions with accelerated levels of genetic differentiation can be detected by generating locus specific or sliding window estimates of sequence dissimilarity when comparing species (for example: Sackton et al., 2019). In contrast, when comparing populations (i.e. when searching for selection on ancestral alleles), genomic regions with accelerated genetic differentiation can be detected by generating locus specific or

sliding window estimates of allele frequency differences, most commonly measured with a Fst metric (Akey et al., 2002; Beaumont, 2005; Wolf and Ellegren, 2017).

The assumption behind genome wide selection scans on reduced representation libraries is that even though random subsets of SNPs are unlikely to contain some – let alone all – sites under selection, selected loci will be within linkage disequilibrium of the sites within the dataset, and these linked sites will exhibit the signal of selection. Dense catalogues, comprising a sufficient number of snps, are needed to reliably screen the entire genome to acquire statistical power in detecting genomic regions under selection (Storz, 2005). Once genomic regions of interest have been detected, the next step is to determine which functional loci – either coding or regulatory sequences – are located within these regions, as these functional loci are the ones possibly under selection.

Research aim of this thesis

In this PhD thesis I will perform whole genome scans to search for genetic fingerprints of selection in genomes of natural populations of deer species. My main aim is to determine if the genomes of my study populations and study species contain evidence of past and/or ongoing events of natural selection. I will do so by investigating the genetic divergence of:

- introduced South Georgia reindeer and their ancestral Norwegian population, which separated $\sim 10^2$ YBP (*Chapter 2*)
- mainland and UK roe deer populations, which separated $\sim 10^4$ YBP (*Chapter 3*)
- extant roe deer species (the European roe deer and the Siberian roe deer), which separated $\sim 10^6$ YBP (*Chapter 4*)

My deeper, ultimate aim is to provide, through analyses of these particular populations and species, more insight into the relative roles of selection and drift under various demographic and environmental scenarios, and as such contribute to the scientific debate about the neutral theory of molecular evolution.

Although the three genomic datasets presented in this thesis total up to an enormous amount of data, these datasets do not suffice to statistically test hypotheses. Many studies such as these, for a wide range of populations and species,

are needed to eventually perform a meta-analysis and obtain an overall emerging picture. Still, I will be able to test some predictions for my particular datasets.

The roe deer populations and species occurred in environmental similar habitats. Although ecological differences can be subtle, diversifying selection can be expected to have been weak or moderate compared to population/species with more pronounced ecological divergence. Although roe deer are provincial (Baker and Hoelzel, 2012), effective population sizes have been found to be relatively high (~10K, Baker and Hoelzel, 2014), providing potential for the manifestation and detection of signatures of selection (because of relatively low levels of genetic drift). Summarized, diversifying selection could be expected to play a minor role in shaping the genetic divergence of roe deer populations/species, but if it did have a role, we should expect to be able to detect the signatures.

Exactly the opposite is presumably true for the South Georgia reindeer founder populations. Having been introduced in an alien environment, there is reason to believe these populations experienced selective pressures to adapt to this new environment. Low effective population sizes, during and following the founder bottleneck, might however have caused dominant drift, overriding the effect of selection, and furthermore complicating the detection of selected loci which did manage to overcome drift. Hence, even though the South Georgia populations might be expected to have experienced positive selection, it is unclear whether this has left detectable signatures of selection, especially given the short time frames.

Outline of this thesis

In this thesis I investigate the extent and causes of genetic divergence of allopatric sister populations on three different time scales. I will discuss each of the three datasets in ascending order of their TMRCA.

In Chapter 2 I study genetic divergence over a time span of 10^2 years (~20 generations). This study centres around SNP datasets from two reindeer (*R. tarandus*) founder populations which were established at the start of the 20th century. Apart from an investigation of genetic divergence on a short time scale, this Chapter focuses on the interplay between selection and drift in founder populations, with the main research question being whether selection can overcome drift in heavily bottlenecked populations.

In Chapter 3 I study genetic divergence over a time span of 10^4 years (~2,000 generations). This study centres around SNP datasets of three western European roe deer (*C. capreolus*) populations. Two of those populations occur on the mainland, whereas a third population represents the native British population, which got cut off from the mainland during the flooding of Doggerland, around $6{\cdot}10^3$ ya. An additional dataset from a fourth population, derived from a recently established founder population in East-Anglia (UK), has been included to provide deeper insight in the relation between TMRCA, population size and genetic divergence.

In Chapter 4 I study genetic divergence over a time span of 10^6 years (~200,000 generations). Whereas in the previous Chapters I investigate differences between populations within the same species, in this Chapter I study differences between different species. I apply a comparative genomics approach by comparing a whole genome sequence of the western roe deer (*C. capreolus*) with a whole genome sequence of its sister species, the eastern roe deer (*C. pygargus*).

In Chapter 5, the general discussion, I will reflect how the findings of the data Chapters reflect on the neutral theory of molecular evolution.

In this thesis I will execute a wide variety of analyses. As the type of the analyses depends on the nature of the dataset, the analyses will differ to a certain extant between thesis Chapters. Whole genome sequences, which are analysed in Chapter 4, allow for example for synteny and gene analyses, which is not possible with SNP data (Chapter 2 and Chapter 3). Still, there are two consistent threads among all three data Chapters:

- For each dataset (i.e. in each Chapter), I will calculate the sequence dissimilarity between sister populations (or sister species).
- For each dataset (i.e. in each Chapter), I will search for genetic differences which are caused by selection, in order to obtain an estimate of the proportion of selective driven genetic differences (alpha).

The analyses applied in all Chapters are generally consistent and can be broadly divided in three categories. The first group of analyses investigate modern and historical population demography and address questions about TMRCA, population sizes and gene flow between the sister populations. The second group of analyses investigates the extent of genetic divergence between the sister populations. This is

expressed in population differentiation estimates (i.e. Fst and Nei's genetic distance) when using SNP datasets, and sequence (dis)similarity indices when using whole genome sequences. The third group of analyses aims to detect which genetic differences result from natural selection as opposed to solely from stochastic forces.

The third and last group of analyses therefore addresses the main question of my thesis: do the genomes of my study populations and study species contain signatures of past and/or ongoing events of natural selection?

Chapter 2

Genetic evidence for parallel insular evolution in the South Georgia reindeer founder populations

Abstract

Founder populations are of special interest to both evolutionary and conservation biologists, but the detection of genetic signals of selection in these populations is challenging due to their demographic history. Geographically separated founder populations subjected to the same selection pressures provide an ideal but rare opportunity to overcome these challenges. I generated an 80K SNP database of two parallel deer founder populations and screened this dataset for signatures of soft sweeps. I find evidence for two genomic regions under selection shared among both populations. I support my findings with Wright-Fisher model simulations to assess the power and specificity of interpopulation selection scans – i.e. Bayescan, OutFlank, PCadapt and a custom-built tool called GWDS – in the context of founder populations. My simulations indicate that loci under positive selection in non-communicating sister founder populations are most confidently detected by GWDS, and provide evidence that the observed outlier regions are true loci under selection. In conclusion, I report a novel selection scan and present empirical evidence for positive selection overcoming drift in heavily bottlenecked founder populations.

Related peer-reviewed publication:

De Jong, M.J., Lovatt, F., Hoelzel, A.R., under review by Molecular Ecology, *Detecting genetic signals of selection in heavily bottlenecked founder reindeer populations by comparing parallel founder events*

Author contributions:

ARH conceived the study and MdJ & ARH wrote the paper. MdJ undertook data, simulation and lab analyses, and developed the selection scan GWDS. FL provided field work and some of the DNA extractions.

Introduction

One of the major current challenges of population geneticists is to discriminate loci under selection from the backdrop of neutral genetic variation (Beaumont, 2005; Oleksyk et al., 2010). For founder populations, which are of special interest to both conservation biologists (Allendorf and Lundquist, 2003) and evolutionary biologists (Templeton, 2008), loci under selection are especially hard to detect due to their demographic history. Genetic drift during and following a founder bottleneck spreads out the backdrop of neutral variation, obscuring the typically weak signals of incomplete selective sweeps (Hermisson and Pennings, 2005) and elevating the false negative and positive rates of selection scans.

Although empirical evidence for adaptation to novel environmental conditions on short, observable time-scales has accumulated in past decades (*reviews on contemporary evolution:* Carroll et al., 2007; Endler, 1986; Hendry and Kinnison, 1999; Reznick and Ghalambor, 2001; Schoener, 2011; *famous examples:* Hof et al., 2016; Johnston and Selander, 1964; Lamichhaney et al., 2015; Reznick and Ghalambor, 2001; *climate change adaptation studies:* Bradshaw and Holzapfel, 2010; Brakefield and de Jong, 2011; Karell et al., 2011; Schilthuizen, 2018), evidence for adaptation specifically in founder populations has so far remained elusive (Colautti and Lau, 2015; Vandepitte et al., 2014). The rarity of empirical evidence for selection in founder populations will in part reflect adaptive constraints of founder populations (Willi et al., 2006), but also the difficulty of detecting the (genetic) signatures of selection within founder populations.

A unique but rare opportunity to overcome the challenges associated with selection analysis in founder populations arises when two or more sister populations (i.e. populations deriving from the same ancestral population) are independently founded in environmentally similar sites (Lee and Coop, 2019). I capitalized on such a system by searching for evidence of selection in two parallel founder deer populations. These founder populations originated in the early 20th century (1911 and 1925), when two small (≤ 10) herds of reindeer (*Rangifer tarandus*) were shipped from Filefjell, Norway, to two peninsula separated by a glacier (Leader-Williams, 1988, page 43) on the island of South Georgia in the South Atlantic Ocean (Leader-Williams, 1980). Despite facing an environment which

differed from their native grounds, both populations established successfully until their cull in 2013 (Leader-Williams, 1988).

As is true for invasive species in general (Allendorf and Lundquist, 2003; Colautti and Lau, 2015), it is not known whether the success of the South Georgia reindeer was aided by adaptation to their novel environment. I reasoned however that if the South Georgia reindeer populations did adapt to their novel environment, the two founder populations potentially underwent parallel evolution. Both populations experienced the same environmental conditions and therefore were subjected to similar selective pressures. If during the founder bottleneck they preserved shared adaptive alleles, this could lead to shared genetic signals of selection. The South Georgia reindeer populations were separated by a glacier (Leader-Williams, 1988, page 43), and therefore shared signals of selection could not have established through gene flow. Since shared signals of selection are easier to distinguish from the background of neutral variation than adaptive loci selected in single populations, the South Georgia reindeer populations provide a promising study system for the detection of genomic signals of selection within founder populations.

I generated an 80K SNP database for both founder populations as well as their common source population, and searched for genetic signals of selection in both founder populations using interpopulation selection scans (Oleksyk et al., 2010). I made use of published selection scans (i.e. Bayescan, OutFlank and PCadapt) as well as of a custom-built tool which I named GWDS, an acronym for Genome Wide Differentiation Scan.

I evaluated the empirical findings by running simulations using a Wright-Fisher model. The main purpose of these simulations was to estimate the probability that the loci marked as outliers by the selection scans were true loci under selection rather than false positives. I did so by assessing the power and specificity of selection scans, including GWDS, in the context of study populations, and specifically in the context of the demographic history of my study populations.

Methods

Library Construction. I selected 120 reindeer samples from an existing DNA archive (Lovatt and Hoelzel, 2014), evenly divided over both South Georgia founder populations and their Norwegian source population. DNA samples were selected based on Qubit quantification scores and molecular weight of the DNA assessed by gel electrophoresis. I constructed two sequencing libraries each of 60 samples following the ddRADseq protocol (Peterson et al., 2012).

Following in silico simulations with the R package SimRAD (Lepais and Weir, 2014), I decided to use a 6 bp cutter (*HindIII*: AAGCTT) and a 4 bp cutter (*MspI*: CCGG), with a fragment size selection window of 250 bp width (by including all fragments with a length of 275 to 525 bp, excluding the adapters), targetting 120,000 loci with an average read depth of 30. By multiplying this expected number of loci against with their average length (250 bp), as well as with a conservative estimate for nucleotide diversity (1/2000), and with an approximation for the harmonic number of Watterson's estimator (Watterson, 1975), I estimated that this size selection window would yield ~50,000 SNPs with a minor allele frequency (MAF) >= 0.05.

The actual size selection was executed with a Sage Science PippinPrep machine. The Phusion High-Fidelity kit was used for a 13 cycle PCR (denaturation step: 62°C for 20sec; annealing step: 72°C for 45 sec; extension step: 72°C for 5 min). Libraries were paired-end sequenced on an Illumina HiSeq_2500 (version 4 chemistry) machine.

SNP calling and filtering. Reads were trimmed to 110 bp and demultiplexed and filtered using STACKS1.35 (Catchen et al., 2013). Unpaired reads were discarded. Paired reads were aligned using the very-sensitive mode of Bowtie version 2.2.5 (Langmead and Salzberg, 2012), against both the reindeer (*Rangifer tarandus*) genome (Li et al., 2017) as well as the cow (*Bos taurus*) genome (Zimin et al., 2009) – cow being at the time the species closest to reindeer with a genome assembly up to the chromosome level. Samtools version 1.3.3 (Li et al., 2009) was used to filter out reads which aligned disconcordantly, reads with a mapping quality below 3, as well as reads which aligned to more than one location in the genome.

SNPs were called using the STACKS refmap pipeline with default settings. Loci for which at least 30 percent of all individuals had a read depth below 8 were removed. I accepted multiple SNPs per read (i.e. I did not set the –write-single-SNPs flag when running the 'populations'-command), as I opted to optionally 'thin' the datasets downstream.

PGDSpider (Lischer and Excoffier, 2012) and PLINK v1.90 (Purcell et al., 2007) were used to convert the output from genepop or vcf format to a genlight object, supported by the R package Adegenet (Jombart, 2008; Jombart and Ahmed, 2011). All samples with more than 25 percent missing data were removed. I discarded SNPs which met any of the following criteria : 1.) >10% missing data (after removal of low quality individuals); 2.) minor allele count (MAC) = 1; 3.) excessive heterozygosity excess (he > (2pq + ½q)); and 4.) excessive read depth. I also filtered out a few SNPs which mapped to the same site of the reindeer genome and yet belonged to different STACKS loci. I optionally thinned the data by selecting one SNP per 500 bp region.

Structure and diversity analyses. For population genetic analyses, I used a filtered and thinned dataset derived from alignment to the reindeer genome. Linkage disequilibrium analysis was executed on reduced datasets excluding SNPs with MAC<5. For selection analyses, I used a filtered, non thinned dataset derived from alignment to the reindeer genome. For genome wide genetic analyses, I used a filtered, non-thinned dataset derived from alignment to the cow genome.

All population structure analyses (PCA, DAPC, admixture analyses) were executed in R, using functions implemented in the Adegenet, Ape (Paradis and Schliep, 2019; Paradis et al., 2004), and LEA (Frichot and François, 2015) packages. For DAPC (run in Adegenet) I set the number of PCs to 1/3th the number of individuals, the number of clusters equal to the number of populations in the dataset (i.e. 3), and the number of discriminant functions to 3.

For admixture analysis in LEA I set K (number of populations) to 2-6, alpha to 10, tolerance to 0.00001, and the number of iterations to 200. To quantify population differentiation I calculated Nei's D (Nei, 1972) using a function implemented in StAMPP (Pembleton et al., 2013), as well as Weir & Cockerham's F_{ST} (Weir and Cockerham, 1984). I assessed genetic diversity by generating site

frequency spectra, MAF histograms, and estimates of sample genome wide heterozygosity.

To estimate sample genome wide heterozygosity, I first determined 'N_seg', the number of segregating sites within the population to which the individual belonged. Second, I calculated for each individual 'He_seg', the proportion of those segregating sites being heterozygous. Finally, I calculated genomeHe using the formula: genomeHe = (He_seg * N_seg)/N_total, in which N_total equals the combined length of all loci/stacks which passed filter settings. This provides an estimates of the proportion of heterozygous sites across all sequences sites, which is a proxy of genome wide heterozygosity.

LD analyses were executed by calculating squared correlation coefficient estimates for unphased data using the software PLINK. I generated LD estimates for all SNP pairs occurring on the same contig at maximum 5Mb apart. Contemporary gene flow was estimated using BayesAss3-SNPs (Mussmann et al., 2019). The number of iterations was set to 1,000,000, burn-in to 100,000, seed to 10, and delta values to 0.1.

Selection analyses. **I** screened the SNP dataset for loci under selection using two approaches: a pooled approach and an independent approach. In the pooled approach I pooled the data of both founder populations and executed selection scans by contrasting the source population to the pooled founder populations (i.e. 'Norway vs Busen & Barff'). In the independent approach I executed selection tests for both founder populations independently by running pairwise comparisons between source and founder ('Norway vs Busen' and 'Norway vs Barff').

To identify positively selected loci, I used a custom-built tool (GWDS; discussed below) as well as three published selection scans: Bayescan (Foll and Gaggiotti, 2008), OutFLANK (Lotterhos and Whitlock, 2015), and PCadapt (Duforet-Frebourg et al., 2014; Luu et al., 2017).

Bayescan is a F_{ST} outlier test which simulates a null distribution of locus specific F_{ST} values and subsequently detects loci which stand out from this simulated distribution. Although not specifically designed for pairwise population comparisons, it could be argued that it is better suited to detect loci under selection in pairwise population comparisons than in study systems consisting of more than

two populations. The reason is that Bayescan implements a method which assumes that all populations are equally related. If there is only a single pair of populations in the data, there is no possibility that one pair of populations is more related than another pair (Lotterhos and Whitlock, 2015), and therefore no possibility that this assumption of Bayescan is violated.

OutFLANK is also a F_{ST} outlier test, but unlike Bayescan it infers the distribution of F_{ST} values from the observed data (rather than simulating it). The software trims from the observed data the top 5% and bottom 5% F_{ST} values (or any other user defined, non-default percentage), fits a chi-squared distribution to the remaining data (the 'core' or 'trimmed' distribution), and subsequently uses the inferred distribution to calculate the right hand sided p-value for each locus (Lotterhos and Whitlock, 2015).

PCadapt is not a F_{ST} outlier test and is based on individuals rather than on populations. Whereas Bayescan and OutFLANK require samples to be a priori assigned to populations, PCadapt infers population clustering from the data by principal component analyses. Subsequently the software regresses each SNP to the principal components, and standardizes the obtained regression coefficients to z-scores. To find outlier SNPs, the obtained vectors of z-scores are translated into Mahalanobis distances. These Mahalanobis distances are subsequently assigned p-values assuming they are chi-squared distributed. The underlying reasoning of PCadapt is that outlier loci suggest aberrant population clustering and therefore have fit less with the principal components than neutral loci. (For more information, see: Duforet-Frebourg et al., 2014; Luu et al., 2017).

Bayescan's false discovery rate (FDR) was set to 0.01, and focus on outlier loci with positive alpha scores, with are indicative of positive/diversifying selection rather than on of balancing/purifying selection. OutFlank outliers were scored based on Holm corrected p-values rather than on q-values, which is the default setting. PCadapt outliers were scored based on Bonferroni corrected p-values, with K set to 2. My simulations (discussed below) showed that above settings resulted in optimal combinations of power and sensitivity. For both the empirical and the simulated datasets, and for both the pairwise and the pooled approach, I ran PCadapt with K equals 2.

I visually assessed the outlier loci by comparing the locus specific Weir and Cockerham He and F_{ST} estimates (Weir and Cockerham, 1984) of outlier loci to those of remaining loci in a He-F_{ST} plot (Beaumont and Nichols, 1996).

GWDS. Next to using published selection tests I also developed a new tool to search for loci under positive selection: GWDS or Genome Wide Differentiation Scan. Similar to GWAS, GWDS compares allele frequencies on a SNP by SNP basis. GWDS differs however from GWAS in two ways. The first difference is that GWDS searches for locus specific associations between allele frequencies and population division, rather than for locus specific associations between allele frequencies and phenotypic traits. This could concern pairwise comparisons between population pairs, but more statistical power is provided by comparisons between sets of multiple populations subjected to contrasting environmental pressures.

The major assumption behind GWDS is that whereas allele frequencies of neutral loci differ randomly among populations, selection will temporarily cause the allele frequencies of positively selected loci to differ more strongly. This reasoning especially holds if both populations are sister populations (i.e. derived from the same ancestral populations), since their allele frequencies are initially correlated. A requirement is that the TMCRA of the sister populations is considerably less than 4*Ne generations, as greater split times will result in fixation of alleles through drift.

Allele frequency differences are scored as p-values outputted by Fisher exact tests executed on contingency tables of allele counts. These p-values are calculated using R's built in fisher.test function, which outputs p-values which are up to 4 decimals identical to p-values outputted by PLINK's fisher's exact test.

The second difference between GWDS and GWAS – as well as between GWDS and methods applied in Hendrickson (2013), Cammen et al. (2015), Shultz et al. (2016) – is in the method of outlier detection. GWDS considers the p-values in itself to be uninformative, as those values depend on sample size, the demographic history (i.e. Ne) of the populations, and the relatedness and connectiveness of the populations. Instead, GWDS searches for loci with p-values which stand out from the overall distribution of test scores. It does so by calculating a Bonferroni corrected right tail value.

GWDS assumes that the distribution of the negative natural log of obtained p-values fits an exponential distribution, and therefore that right tail values can be derived from the observed mean (i.e.: rate = 1/mean). This assumption of GWDS holds if and only if both sister populations split less than 4·Ne generations ago. Longer split times will result in a bimodal distribution, with the modi reflecting fixation and loss of alleles due to drift. Another assumption is that the vast majority of SNPs will be neutral, and that the observed mean will not be inflated by a few loci under selection.

The right tail threshold p-value is subsequently calculated for an Bonferroni corrected alpha value (i.e. alpha/n_snps, with alpha set to 0.05). SNPs with a –ln(p-value) greater than this right tail value are marked by GWDS as outliers, possibly representing loci under positive selection.

Although GWDS has been developed to detect soft sweeps (i.e. selection on standing variation, Hermisson and Pennings, 2005, 2017), it has the potential to detect selection on new mutations as well, provided the migration rate between populations does not equal zero. In the absence of gene flow, genetic drift will ultimately cause fixation or loss of neutral alleles in both populations, resulting in a bimodal distribution of Fisher's exact test scores. GWDS will not be able to fit an exponential distribution to this bimodal distribution, and likely return either zero outliers or high false positive rates. GWDS is therefore applicable only to recently diverged isolated sister populations, or to ancient sister populations with correlated allele frequencies due to gene flow.

As can be inferred from the explanation on the inner workings of GWDS, several key differences exist between GWDS and existing selection scans. These differences have consequences for the applicability of the method, and are therefore worth to mention explicitly:

- One key difference between GWDS and F_{ST} outlier tests is that GWDS measures the differentiation of loci between populations using p-values outputted by rfisher exact test on allele counts contingency tables. These p-values are a measure of differences in relative proportions of minor and major alleles, and not a measure of the likelihood that a locus is an outlier, and a measure of the likelihood that a locus is an outlier.

- A second key difference between GWDS and F_{ST} outlier tests (except OutFLANK) is that GWDS does not simulate a null distribution. Instead GWDS fits a probability distribution to the observed dataset, similar to OutFLANK. Unlike OutFLANK, GWDS fits an exponential distribution to (the negative log of) rfisher exact test p-values (rather than a chi-squared distribution to F_{ST} values) and does not trim the dataset (i.e. no removal of top 5% and bottom 5% values). Furthermore, whereas OutFLANK uses the obtained probability distribution to assign right tail p-values to loci, GWDS uses the obtained probability distribution to determine the right tail threshold value. Because GWDS does not attempt to simulate a null distribution, changes in population size through time (e.g. population expansions) do not compromise the power and specificity of GWDS, as has been reported for several existing selection scans, including Bayescan (i.e. figure 3 in Luu et al., 2017).
- A third key difference between GWDS and F_{ST} outlier tests (except OutFLANK) is the underlying demographic model. Originally F_{ST} outlier tests simulated the null distribution of F_{ST} values assuming an island model consisting of an infinite number of equally sized populations ('demes') with equal migration rates between populations and with no hierarchial structure (i.e. all population pairs are equally related) (Beaumont and Nichols, 1996). New, more sophisticated methods, such as the one implemented in Bayescan, liberated F_{ST} outlier tests from the first two constraints and allowed the study system to contain a limited number of populations of various sizes and with unequal migration rates among pairs of populations (Beaumont and Balding, 2004; Foll and Gaggiotti, 2008). The underlying demographic model of Fst outlier tests has remained however unchanged and still comprises an island model consisting of several (two or more) populations which are derived from a common ancestral gene pool and which potentially exchange migrants. The underlying demographic model of GWDS, in contrast, is limited to the specific case of two (either equal or unequally sized) populations which are derived from a common ancestral gene pool (with or without gene flow). In short, GWDS is specifically designed for pairwise comparisons

between populations, whereas existing selection scans, including Bayescan, OutFLANK and PCadapt, can also detect signals of selection in study systems which consists of more than two populations.

Unlinked SNPs simulations tool. Simulated datasets of founder and source populations were generated using custom R functions describing a Wright-Fisher model. The demographic scenario consisted of a source population with a constant Ne of 1000 individuals which buds at t0 a founder population. Both the source and the founder population are subsequently allowed to drift for a certain number of generations. The source and the founder population do not exchange migrants (i.e. no gene flow).

The simulation tool simulates changes in allele frequences through generation of standing variation. It does not incorporate new mutations. The starting allele frequencies in the source population were set to 0.15 and allowed to drift for 200 generations before the founder event. Founder events and genetic drift subsequent and prior to the founder event were simulated with the rbinom function, which outputs the number of successes (number of allele copies in next generation) given a sample size (2·Ne) and a success probability (allele frequency in current generation).

Selection was simulated in two ways. For selection coefficients of s>0.05, I simulated selection as a continuous process by multiplying each generation the rbinom output with the selection factor (1+s). For s<=0.05, I opted for a different way because the effect of selection was counteracted by rounding. Here I simulated a selective event as a doubling of the number of adaptive alleles, with a probability of occurrence of s per generation. Neutral loci were defined as SNPs which allele frequencies were affected by drift only. Loci under selection were defined as SNPs which allele frequencies were affected by both drift and selection. It was assumed that adaptive alleles were minor alleles in the source population, being (nearly) neutral in the source population habitat but advantageous in the founder population habitat. The opposite scenario, in which a (nearly) neutral major allele becomes detrimental in the founder population habitat, would produce the same outcome of high allele frequency differentiation.

The output of the simulations were allele frequencies/counts of a source and two founder populations, both directly following the founder event (t_0) and after a certain number of generations (t_{gen}). To incorporate observer's error related to limited sampled sizes (i.e. deviation between population allele frequencies and observed allele frequencies), I generated sample allele frequencies using the rbinom function, with number of successes representing the number of allele copies in the genotyped individuals given, sample size equalling 30 individuals, and with success probability being defined as the population allele frequency. Sampled t_{gen} output vectors served as simulated input for selection scans.

Validation of unlinked SNPs simulation tool. I validated my Wright-Fisher simulator by comparing three simulation output scores with theoretical expectations: 1.) the proportion of retained variation directly after a founder event; 2.) the fixation probability of neutral and adaptive alleles; 3.) time to fixation. Consistent with expectations, the observed proportion of retained variation depended on the number of founder (Nf) and allele frequencies in the source population as follows: $1-(1-maf)^{2 \cdot Nf}$ (Fig. 2.4A). Fixation probabilities approximated $(1 - e^{-2 \cdot Ne \cdot s \cdot p})/(1 - e^{-2 \cdot Ne \cdot s})$ (Fig 2.1B in Kimura, 1962), which for nearly neutral alleles ($s \dashrightarrow 0$) corresponded to the mean frequency of alleles directly following the founder event, as expected for neutral alleles ($s=0$) (Fig 2.1B). I also confirmed that fixation times were less than 4Ne generations for neutral alleles (Fig. 2.1C) and fixation times of less than $(2/s) \cdot \ln(2 \cdot Ne)$ generations for adaptive alleles (Fig 2.1D, Kimura and Ohta, 1969). This is consistent with expectations, because the fixation time of standing variation should fall below the fixation time of new mutations.

Unlinked SNPs simulation analyses. After validation of my simulator, I used it to assess the performances of GWDS in comparison to PCadapt and OutFLANK. (Bayescan, which runs on Linux rather than in R and has relatively long computation times, was excluded from this part of the analysis.) For each test – GWDS, OutFlank and PCadapt – I calculated the false positive rate (1 – specificity) as the number of neutral SNPs marked as outliers divided by the number of neutral SNPs with MAF > 0 at t_0. Similarly, for each test I calculated the power (1 – false negative rate) as the number of selected SNPs marked as outliers divided by the number of selected SNPs

with MAF > 0 at t_0. (In other words: adaptive loci which were lost during the founder bottleneck event were excluded from the power and specificity calculations.)

To evaluate the best approach for multiple testing correction, I corrected the p-values generated by each test using three correction methods: Benjamini-Hochberg, Bonferroni, and Holm correction. I also evaluated the performance of each test without correcting the p-values, resorting to q-values in the case of OutFLANK (default setting).

I first ran simulations for a range of demographic scenarios, for 9000 neutral loci and 1000 adaptive loci per scenario. Demographic scenarios included all combinations of selection coefficients ranging between 0 and 0.2 (step size 0.025) and constant effective founder population sizes of 10, 20, 30, 50, 100, 200, 300, 500, and 1000. The number of founders was set equal to the effective population size, and the number of generations of the founder population (starting at the founder event) was set to 20.

This first round of simulations was executed to evaluate the power and specificity of the three selection scans (GWDS, Outflank and PCadapt) under various demographic scenarios and using various correction methods for multiple testing (i.e. none (FDR-rate approach in the case of OutFlank), Benjamini-Hochberg, Bonferroni, and Holm method). GWDS and OutFlank were instructed to calculate the neutral distributions based on the neutral loci only. Because this option is not available for PCadapt, and so that PCadapt could reliably obtain a neutral distribution, I set the proportion of adaptive alleles to a maximum of 0.1 (i.e. 1000/10000).

Following the outcome of this initial round of simulations, I ran a final simulation to estimate the power and specificity (and hence false discovery rate) of selection scans given the demographic scenario of the South Georgia reindeer populations. As the neutral distributions depend on the number of loci, I ran this simulation with the same number of loci as the empirical datasets (i.e. 80000 loci). I used an unrealistic number of 1000 adaptive loci to obtain a precise estimate

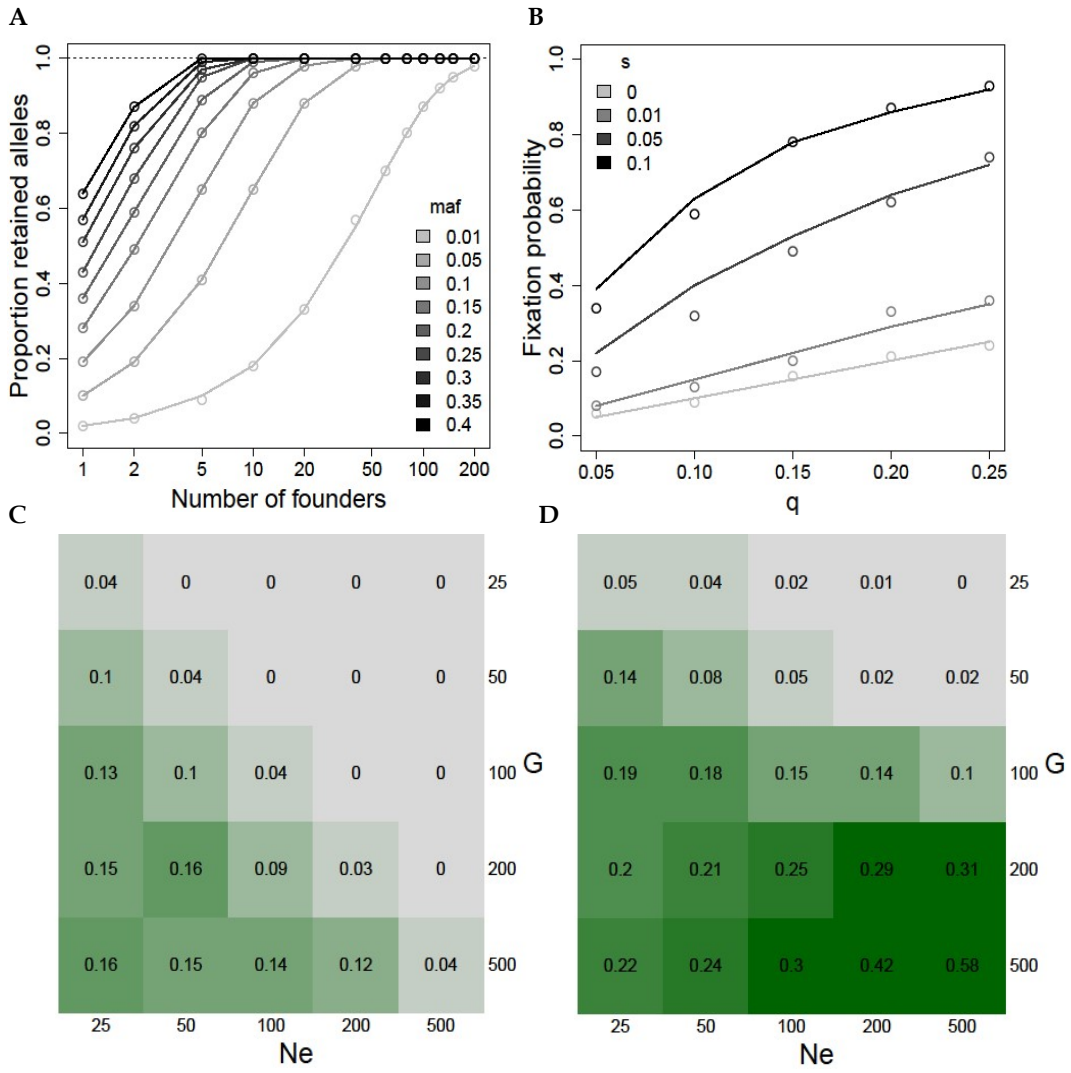


Fig. 2.1. Validation of simulation model. *Retained variation (A) and fixation probabilities (B-D) in founder populations, as inferred from my simulation tool, depending on the number of founders, the effective population size (Ne), the age of the population (number of generations (G), the selection coefficient (s), and the mean of an uniform distribution of minor allele frequency in the source population (maf or q). All estimates are obtained from 1000 SNPs.* ***A. Retained variation.*** *Simulated (points) and expected (lines) retained variation in diploid founder populations directly following the founder event given the number of founders and given the mean allele frequency in the source population.* ***B. Fixation probability****. Simulated (points) and expected (lines) fixation probabilities in diploid founder populations given the selection coefficient and given the mean allele frequency in the source population. Number of generations is set to 500. Ne is set to 50.* ***C. Time to fixation for neutral alleles****. Simulated fixation probabilities in founder populations given a fixed effective population size (Ne = 25, 50, 100, 200, 500) and given the age of the founder population (G = 25, 50, 100, 200, 500). s is set to 0. q is set to 0.15. Theoretical populations genetics predicts a fixation time of a neutral newly mutated allele of 4Ne generations.* ***D. Time to fixation for positively selected alleles.*** *As C, but with s to 0.1. Expected fixation time of adaptive neutral allele equals 2/s * ln (2*Ne) generations.*

of the false negative rate. (Whilst running the selection GWDS and OutFlank, neutral distributions were derived from datasets which excluded the 1000 adaptive loci.) The demographic scenario settings were 10 founders and a constant effective founder population size of 50 individuals during 20 generations. Based on the outcome of the first round of simulations (see results section), I adjusted PCadapt p-values using the Bonferroni correction, and OutFlank p-values using the Holm method. GWDS p-values were not adjusted.

Gene identification. Genes and other genomic features close to outlier SNPs were identified based on both a cow genome annotation (ref Bos_taurus_UMD_3.1.1, Zimin et al., 2009) and a reindeer genome annotation (Li et al., 2017), using the software BEDtools v2.26.0 (Quinlan and Hall, 2010). I decided to use the average spacing between SNPs as the maximum accepted distance between the outlier SNP and gene. Genes were considered potentially linked to an outlier SNP if they were within 150 kb distance from the outlier SNP, and if no non-outlier SNP was present in between.

Results

SNP calling and filtering. Both sequencing lanes combined produced 692.7 million (320.3 + 372.4) single-end reads. Over 13.0 million reads had to be discarded due to either low quality, an ambiguous radtag, or a missing read mate, resulting in an average number of 2.8 million read pairs per sample (stdev: 1.6 million, min: 0.2 million, max: 7.4 million) (Table A2.1, Fig. A2.1). Mean alignment rates equalled 95% and 65% respectively, with 85% and 43% of the reads aligning concordantly to one location (Fig. A2.2).

From the reindeer aligned dataset, STACKS obtained 418,286 loci/stacks, of which 87,552 loci/stacks passed the filters, consisting of 9,627,420 sites, of which 87,876 (0.91%) sites were bi-allelic SNPs – 47,152 of which with MAF >= 0.05 – concentrated on 50,967 loci/stacks. The mean sequence depth per per individual ranged from 0 to 67 reads per locus, with an average of 26 reads per individual (Fig. A2.3). Individuals with less than 1, 0.5 and 0.25 percent missing data had a minimum cover of respectively 26, 32 and 35 reads per locus (Fig. A2.3). I retained 95 individuals (30-34 individuals per population) and 83,406 SNPs after

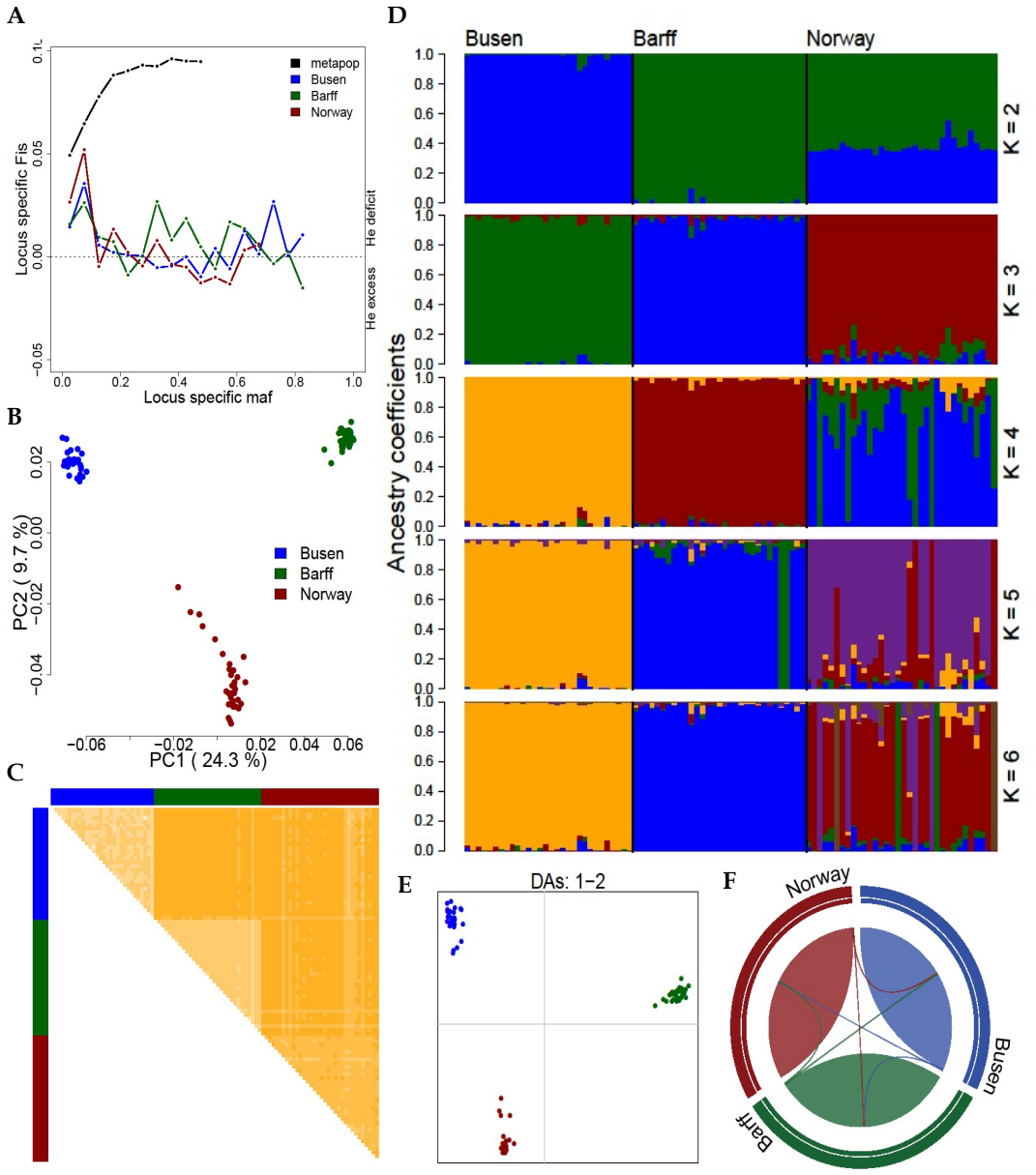


Fig. 2.2. Genetic clustering analyses of reindeer samples belonging to both South Georgia populations (Busen and Barff) and their Norwegian source population. *Colour coding (except for D): blue = Busen, green = Barff, red = Norway.* ***A.*** *Fixation index ((He – Ho)/He) per population, displaying absence of Wahlund effect.* ***B.*** *Principal coordinates analysis based on Nei's genetic distance.* ***C.*** *Nei's genetic distance between samples.* ***D.*** *Admixture analyses for 2 <= K <= 6, with random colour coding.* ***E.*** *Discriminant analysis of principal components.* ***F.*** *Migration rates between the three populations, as inferred by Bayesass3-SNPs.*

filtering and 27,690 SNPs after thinning. The GC-content equalled 0.6, and 'transversion vs transition'-ratio ranged between 1.96 and 2.16, depending on the filter settings (Fig. A2.4, Table A2.2).

From the cow aligned dataset, STACKS obtained 205,076 loci/stacks, of which 36,273 loci/stacks passed the filters ('sample/population constraints'), consisting of 3,990,030 sites (STACKS claims 3,969,066), of which 29,037 (0.72%) sites were bi-allelic. These biallelic sites were concentrated on 18,762 loci/stacks. I retained 95 samples and 20.184 SNPs after filtering and thinning. The SNPs were evenly spread over chromosomes (Fig. A2.5), with a median and average spacing of respectively 0.2 and 23 kb for the filtered dataset and 38 and 60 kb for the thinned dataset (Table A2.2).

Structure and diversity analyses. Population structure analyses (i.e. PCA, DAPC, admixture analyses) verified the existence of three distinct clusters, and therefore the assumption that the two founder populations were geographically isolated (i.e. no gene flow; Fig. 2.2A-E, A2.6). Absence of migration was furthermore confirmed with the software BayesAss3-SNPs (Fig. 2.2F, Table A2.3). Population specific genetic diversity estimates showed strong signatures of recent bottleneck events, with both founder populations displaying site frequency spectra (SFS) typical for bottlenecked populations: reduced nucleotide diversity coupled with high proportions of common SNPs, testifying that many alleles, mostly of low frequency, were lost during and/or after the founder bottlenecks (Fig. 2.3).

Estimates of genome wide proportions of segregating sites equalled 0.43%, 0.49%, and 0.83% respectively for Busen, Barff and Norway (Fig. 2.3E). The equation $1-(1-maf)^{2 \cdot Nf}$ can be used to calculate expected proportions of segregating sites directly following the founder bottleneck. Given that the average MAF within Norway equalled 0.17, and assuming the number of founders (Nf) of the Busen and Barff populations were respectively 7 and 10 individuals (Leader-Williams, 1988), the proportion of segregating sites directly following the bottleneck will have been around respectively 0.77% and 0.81%, much higher than 0.43% and 0.49%. The implication is that the majority of genetic variation was lost due to genetic drift during subsequent generations rather than during the founder event itself.

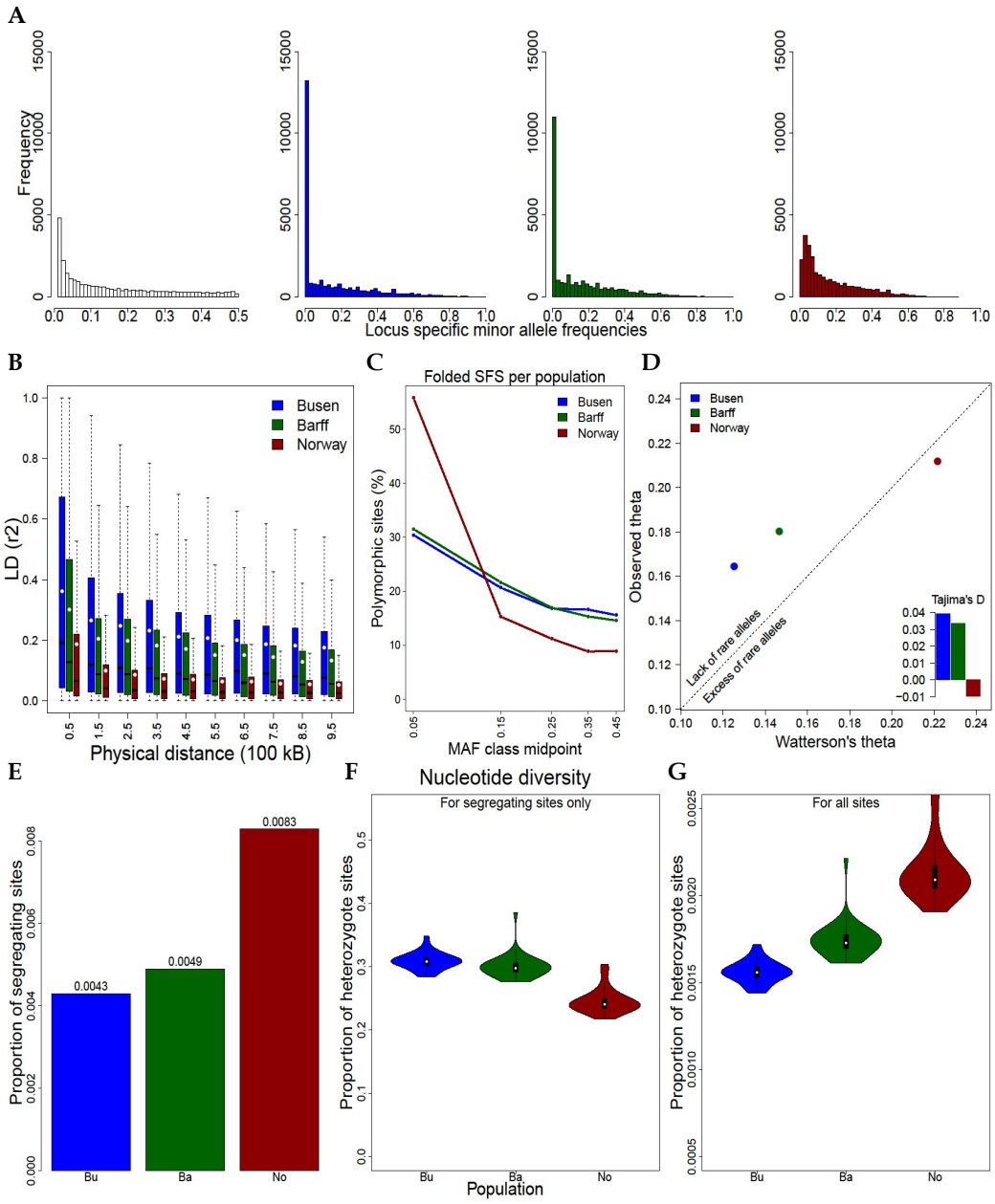


Fig. 2.3. Genetic diversity estimates for both South Georgia reindeer populations (Busen, Barff) and their Norwegian source population. *Colour coding: blue = Busen, green = Barff, red = Norway, white = metapopulation.* ***A.*** *Histograms of minor allele frequencies.* ***B.*** *Boxplots of linkage disequilibrium estimates (squared Pearson correlation coefficients based on genotype scores) per physical distance class (100 kB bins). White dots indicate mean values.* ***C.*** *Percentage of segregating sites per minor allele frequency class.* ***D.*** *Observed theta versus Watterson's estimate of theta. Inset: Tajima's D, scaled to 1bp.* ***E.*** *Proportion of segregating sites over all sequenced sites.* ***F.*** *Estimates of sample heterozygosity, obtained by only considering sites which are segregating in the population to which the sample belongs.* ***G.*** *Estimates of sample heterozygosity, over all sequenced sites.*

Similar conclusions can be drawn from heterozygosity estimates. Mean locus heterozygosity estimates per population, when considering both segregating and non-segregating sites), were 0.19, 0.17 and 0.24 (Fig. A2.7) for Barff, Busen and Norway. These relative values (differences between populations) are in agreement with conclusions previously drawn based on microsatellite analyses (Lovatt and Hoelzel, 2014). Expected heterozygosity (He) depends on Ne and He in the previous generation as described by the function: $He_t = 1 - 1/(2 \cdot Ne) \cdot H_{t-1}$ (Nei et al, 1975). Expected mean locus heterozygosity of the Busen and Barff founders therefore equals respectively 0.22 and 0.23, much higher than 0.17 and 0.19. Again the implication is that the majority of genetic variation was lost due to genetic drift during subsequent generations, rather than during the founder event itself.

Selection analyses. The number of outliers identified by Bayescan, GWDS, OutFlank and PCadapt for the pooled approach were respectively 10, 3, 5 and 15 (Fig. 2.4C-D, A2.8A). None of the outliers detected by Bayescan had a negative alpha value (Fig A2.8B). Overlap between the sets of outliers identified by different scans was restricted to three outliers marked by both Bayescan and GWDS, of which one was also identified by PCadapt (Fig 2.4C). Two of those overlapping outliers were, according to alignments to both the cow genome and the reindeer genome, adjacent SNPs 80-85 kB apart.

The two adjacent SNPs mapped to a genomic region of cow chromosome 25 displaying a weak peak-valley-peak signature indicative of positive selection in sister populations: F_{ST} peaks for both source-founder comparisons, and an F_{ST} valley for the founder-founder comparison (Fig 2.4B, A2.9-A2.10, Roesti et al., 2014). As expected for a soft and incomplete selective sweep (Hermisson and Pennings, 2005) sliding window Tajima's D analyses did not reveal a signal of selection for this genomic region (Fig. A2.11).

The population specific MAFs (with the minor allele defined respective to the metapopulation) of the most confidently marked outlier SNP equalled 0.03, 0 and 0.68 for respectively Busen, Barff, and the Norwegian source population. The

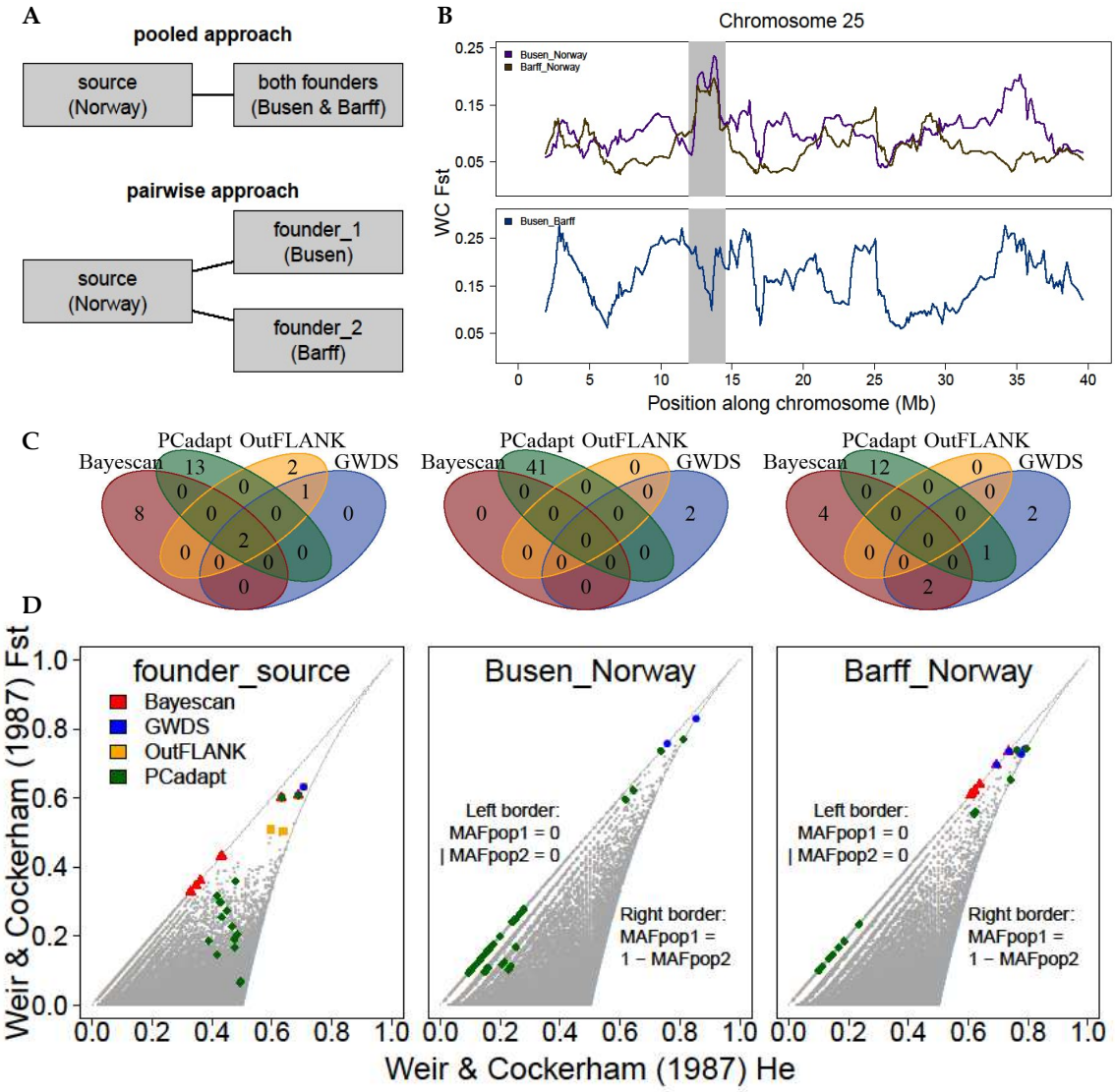


Fig. 2.4. Selection analysis. A. *Conceptual model of the two approaches used when running selection analysis.* ***B.*** *Peak-valley signal around position 14Mb on chromosome 25, the location of the 2 adjacent outlier SNPs.* ***C.*** *Venn diagrams of outlier sets outputted by the selection scans Bayescan, GWDS, PCadapt and OutFlank, for both the pooled approach as well as both comparisons of the pairwise approach.* ***D.*** *Fdist plots showing the location of neutral the outliers outliers outputted by the selection scans Bayescan, GWDS, PCadapt and OutFlank, for the pooled approach as well as both comparisons of the pairwise approach.*

adjacent SNP had MAFs of 0.1, 0.04 and 0.77. The third outlier SNP had a MAF of 0.07, 0.03, and 0.73. Hence, the three outlier SNPs show a consistent signal of positive selection on an allele with a low frequency in the source population.

However, none of the three outliers identified by GWDS in the pooled approach were identified by any of the selection scans for pairwise population comparisons. Overlap between selection scan outlier sets per pairwise comparison was restricted to the Barff-Norway comparison, with one SNP marked as outlier by both GWDS and Bayescan, and two SNPs marked by both GWDS and Bayescan (Fig. 2.4).

He-F_{ST} scores of outlier loci clustered by selection scan (Fig 2.4D). For pairwise comparisons (i.e. Barff-Norway and Busen-Norway), the He-F_{ST} of outlier loci did generally not stand out from the observed overall He-F_{ST} distribution (Fig 2.4D). The opportunity for outlier loci to stand out from neutral loci was limited because the overall He-Fst distribution filled the entire spectrum of possible He-F_{ST} values for pairwise population comparisons. This spectrum of possible He-F_{ST} values has the shape of a shark fin, of which the left boundary is described by F_{ST} = He and represents loci which are segregating in one population only. The right boundary of the 'shark fin'-spectrum represents loci with opposing allele frequencies in either population (e.g 0.3-0.7 in one population and 0.7-0.3 in the other population).

The overall distribution of He-F_{ST} estimates was less inflated for the pooled dataset compared to either pairwise datasets (Fig 2.4D), increasing the opportunity for loci under selection to stand out from the neutral distribution and hence to be detectable by selection scans. Indeed, the outliers detected with the pooled approach (i.e. both founders vs source) did stand out from the overall distribution, except for the majority of outlier loci detected by PCadapt (Fig 2.4D).

Simulation analyses unlinked SNPs. I used the Wright-model simulator for unlinked loci to address several questions about my empirical findings. The main purpose was to assess whether and/or which loci marked as outliers were true loci under selection (questions 4-6). To answer these questions, I required a better understanding of the observed inconsistencies in results obtained from different selection scans and different approaches. The first purpose of my simulations was

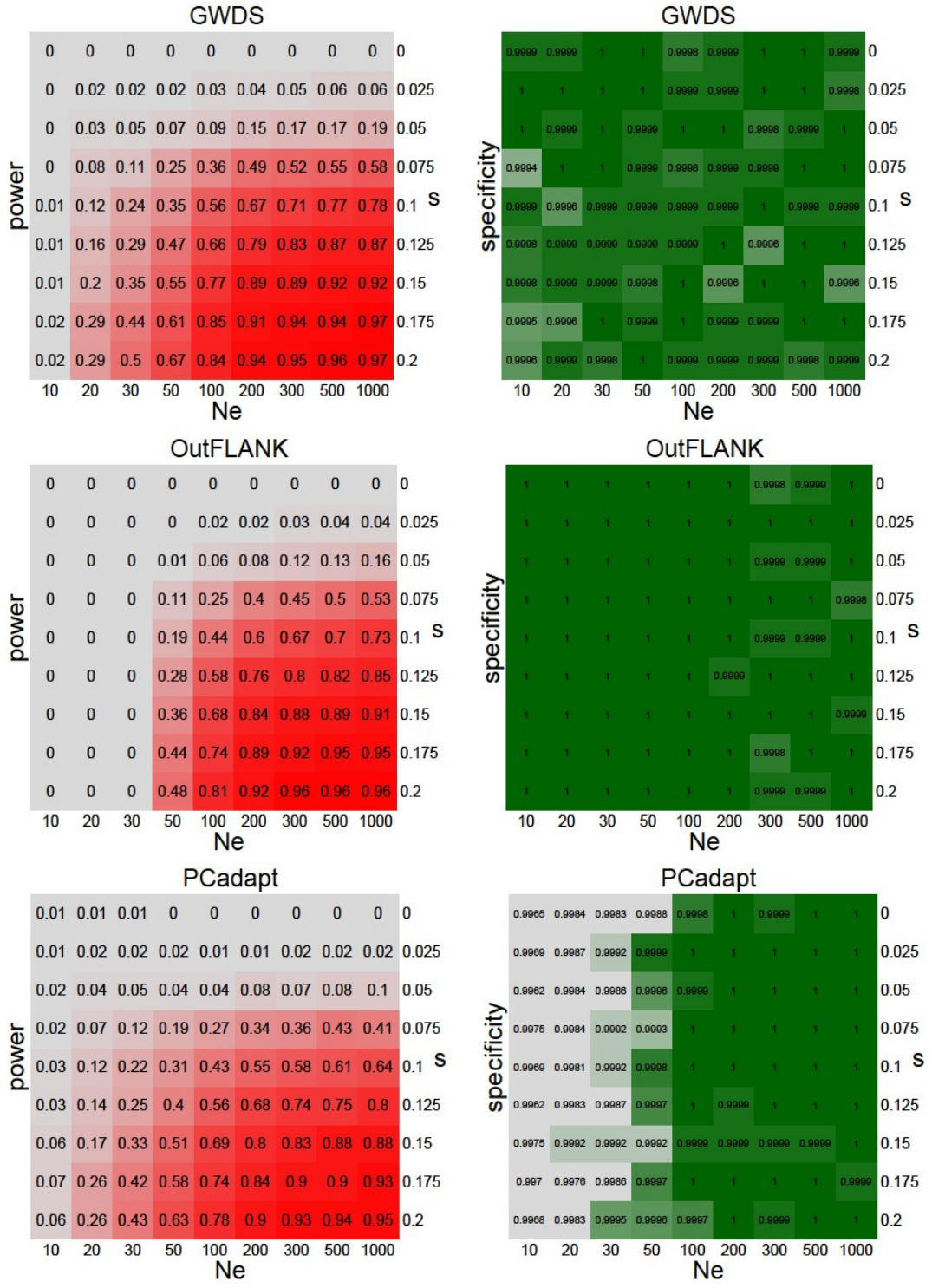


Fig.2.5. Selection scan power analysis. *Power and specificity of the selection scans GWDS, OutFlank and PCadapt in recently established founder populations given a population age of 20 generations,a sample size of 30 individuals per population, a selection coefficient s, and a constant effective population size Ne (i.e. no founder bottleneck). OutFlank and PCadapt p-values were corrected using respectively the Holm and the Bonferroni method. Scores based on 9000 neutral SNPs and 1000 adaptive SNPs.Number of founders equals founder Ne (i.e. no bottleneck).*

therefore to evaluate the performances (i.e. power and specificity) of selection scans, including GWDS, in the context of founder populations (questions 1-3).

The first question I needed to answer in order to be able to compare selection scans, was: Which multiple test correction method maximized the performances of the selection scans used in my simulations? I found that OutFlank and PCadapt generally perform best when using respectively the Holm method and the Bonferroni method for multiple test correction (and hence not q-values, which is the default setting of OutFlank) (A2.12). I also found that the specificity of Bayescan quickly drops when increasing the false discovery rate (FDR), whereas the power of Bayescan only marginally increases (Fig A2.13). I therefore set the FDR of Bayescan to a low value of 0.01, for both simulated and empirical datasets.

My second question was: What is the power and specificity of GWDS under various demographic scenarios of recently established founder populations (TMRCA <= 20 generations), and how do these test scores compare to the power and specificity of other selection scans, more specifically OutFlank and PCadapt? I found that GWDS generally has higher specificity scores (i.e. lower false positive rates) than PCadapt, and higher power scores (i.e. lower false negative rates) than OutFlank (Fig 2.5). This is especially true for scenarios involving relatively low effective population sizes (Ne < 50), as low Ne negatively affects the power of OutFlank and negatively affects the specificity of PCadapt (Fig 2.5). Each of the three selection scans (GWDS, OutFlank, PCadapt) failed to detect the majority of positively selected loci in founder populations which are founded recently (<20 generations ago) and which are small to moderate in size (Ne<100) (Fig 2.5).

My third question was: to what extent do these high false negative rates of selection scans in small founder populations reflect poor test design, and to what extent the outcome of drift overriding and obscuring positive selection? In other words: to what extent do loci under selection stand out from the backdrop of variation found within neutral loci? I addressed this question by visual inspection of the He-F_{ST} distribution of neutral loci.

Simulated He-F_{ST} plots indicated that whether selected loci stand out from neutral loci, depends both on the sample size (number of genotyped individuals

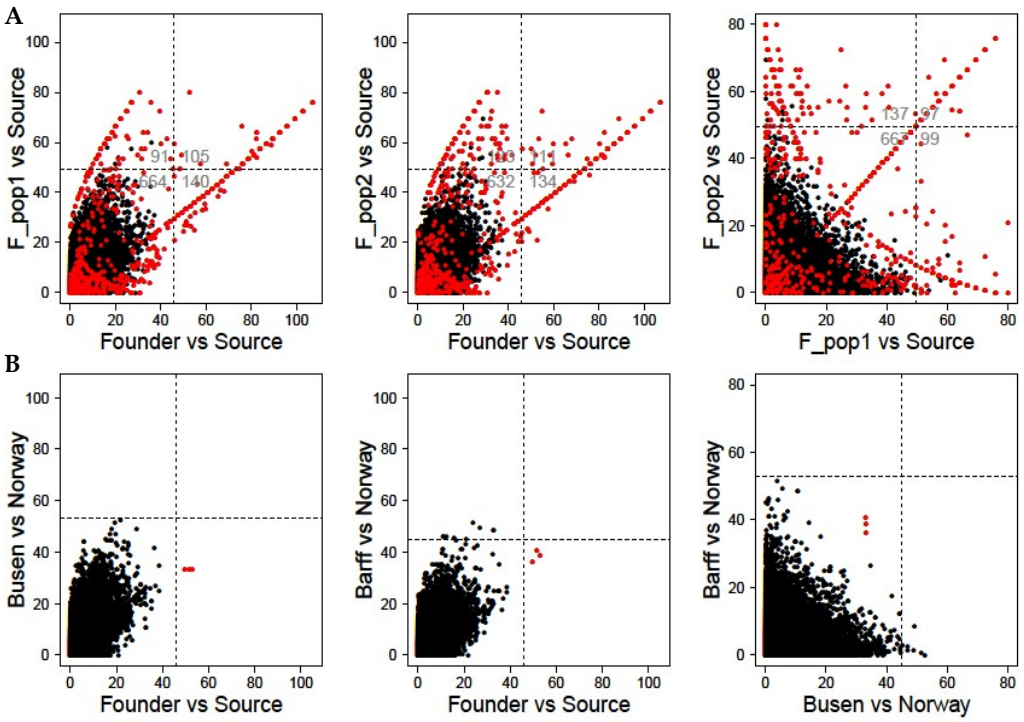


Fig.2.6A-B. Overlap between approaches. *Scatterplots showing the overlap between outliers scored using different approaches (pooled vs pairwise approach) of both simulated (A, upper row) and empirical (B, lower row) datasets. All simulations are based on 79000 neutral loci and 1000 loci under selection (s=0.1), and a two-step demographic scenario consisting of a bottleneck of 10 individuals for 1 generation,and a fixed Ne of 50 individuals during 20 subsequent generations.* ***A.*** *Scatterplots comparing simulated -log10(p-values) of Fisher exact tests performed on contingencies tables of minor allele counts) for simulated datasets using different approaches. First plot: pooled (Founder vs Source) vs pairwise (F_pop2 vs Source). Second plot: pairwise1 (F_pop1 vs Source) vs pairwise2 (F_pop2 vs Source).* ***B.*** *Idem as D, but for empirical datasets. The pooled approach is denoted as 'Founder vs Source'.*

per population) and on the long-term effective population size (Ne) of the founder population (Fig A2.14). For Ne <= 20, the distribution of neutral alleles fills the entire shark fin shaped He-F_{ST} spectrum, obscuring all loci under selection. For Ne >= 50, the distribution does not fill the entire He-F_{ST} spectrum (Fig A2.14). This provides the opportunity for loci under selection to stand out from neutral loci, and therefore to be detectable by selection scans (Fig A2.14).

I furthermore observed that in small founder populations (e.g. Ne = 20), in which drift is dominant, the selected loci have a bimodal distribution on the line He=F_{ST} (Fig A2.14). The group with low He and F_{ST} scores represent loci which were lost in the founder population after the founder event, due to genetic drift. The group with high He and F_{ST} scores represent loci which reached fixation in the founder population. The proportion of selected loci belonging to the first group decreases with increasing Ne (Fig A2.14).

My fourth question was: which demographic model fits the demographic history of the Busen and Barff populations? The answer to this question was needed in order to be able to address the remaining questions. I inferred this model visually by comparing the fit between observed (Fig 2.4) and simulated (Fig. A2.14) distributions (under various demographic scenarios) of locus specific He-F_{ST} estimates, as well as between observed and simulated distributions of GWDS scores (Fig. A2.15B). From these comparisons, I inferred that the demographic history of the Busen and the Barff populations can be roughly described by a two-step demographic scenario, consisting of a bottleneck of 10 effective founders for 1 generation, and a fixed Ne of 50 individuals during 20 subsequent generations.

My fifth question was: Given the demographic history of the South Georgia reindeer populations, which approach (i.e. pooled or independent/pairwise approach) maximizes the performance of selection scans? I found that for the demographic scenario described above, selection scans scored both higher power (Fig A2.15) and specificity (Fig. 2.6A) with the pooled approach compared to the pairwise approach. I also observed that the majority of adaptive loci which were marked as outliers with the pooled approach were not marked as outlier with the independent approach, and vice versa (Fig. 2.6A).

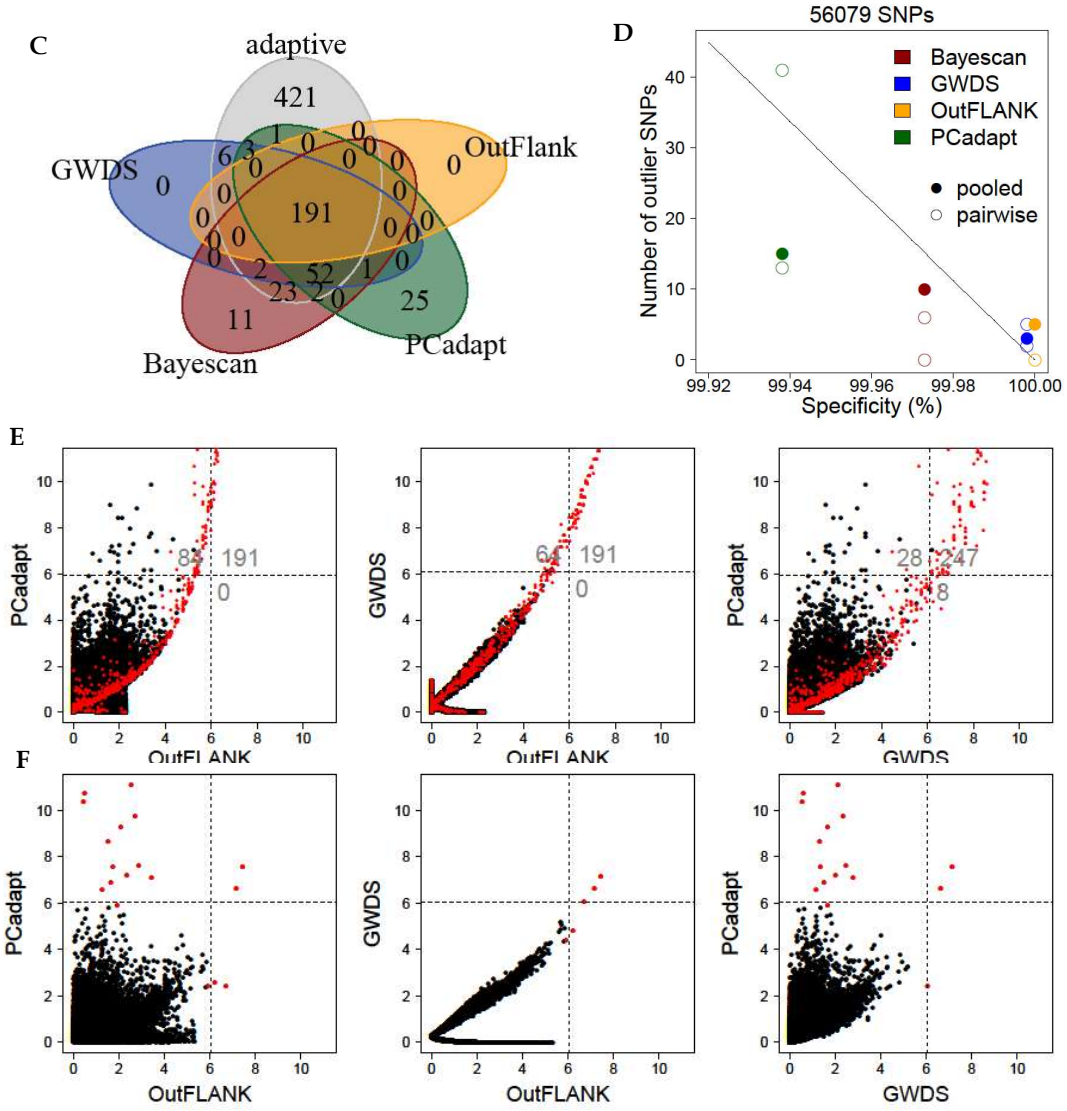


Fig.2.6C-F. Overlap between selection scans. *Overlap between outliers scored by the selection scans Bayescan, GWDS, OutFLANK and PCadapt in simulated (C,E) and empirical (F) datasets. Simulations are based on 59000 neutral loci and 1000 adaptive loci (s=0.1), and a two-step demographic scenario meant to reflect historical Ne of both St Georgia reindeer populations: a bottleneck of 10 individuals for 1 generation,and a fixed Ne of 50 individuals during 20 subsequent generations.* ***C****. Venn diagram showing the simulated overlap between outlier sets and true loci under positive selection.* ***D****. Expected number of false positives (black line), calculated as (1-specificity)*56079 SNPs, versus the number of putative outliers outputted by selection scans for all three comparisons (i.e. Barff vs Norway, Busen vs Norway, and Barff & Busen vs source. Specificity estimates were calculated from simulated data (see 2.6C) using the formula (1-false positives)/79000. The estimate for GWDS was lowered from 100% to 99.95% based on results presented in Fig. 2.5.* ***E.*** *Scatterplots comparing negative log(p-values) of selection scans for simulated neutral (black) and positively selected (red) SNPs using the pooled approach. Dashed lines indicate Bonferroni threshold for 60K SNPs. .* ***F.*** *Idem as E, but for empirical rather than simulated datasets. Red dots indicate SNPs marked by the selection scans as outliers.*

Furthermore, in a 2D-Manhattan plot displaying GWDS test scores for both independent pairwise comparisons, the three outlier SNPs were positioned in a plot region which according to my simulations holds adaptive loci exclusively (Fig 2.6B).

My sixth and final question was: Given the demographic history of the study populations, what is the probability that the outliers detected by the selection scans are true loci under selection? When applying the pooled approach to a simulated dataset generated with the demographic scenario described above, Bayescan, GWDS, OutFLANK and PCadapt marked respectively 18, 0, 0 and 37 out of 79000 neutral loci as false positives, translating to specificity scores of respectively 99.98%, 100%, 100% and 99.96% (Fig 2.6C). The total number of outlier SNPs marked by the four selection scans in my empirical datasets fit the expected number of false positives based on these specificity scores and the size of my dataset (Fig 2.6D), suggesting that all outlier SNPs could represent false positives. I however also found that nearly all SNPs detected by more than two outlier scans were true adaptive loci (Fig 2.7E), suggesting that the three outlier loci detected by multiple selection scans (Fig 2.7F), were true loci under selection.

Gene identification. As mentioned above, among the three identified outlier SNPs two were on the same contig and the third ('non-adjacent') SNP was on a different contig. The closest known gene to the non-adjacent outlier SNP is HAO1, which codes for the protein hydroxyacid oxidase. This gene is however separated from the outlier SNP by a stretch of 200kB containing four non-outlier SNPs, and is therefore unlikely to be of interest (Fig. 2.7).

In contrast, I did find a gene relatively close to the two adjacent outlier loci. Alignments to both the reindeer and the cow genome indicated the presence of an exon in between the two adjacent outlier SNPs (Fig. 2.7). This exon is part of a gene coding for myocardin-related transcription factor B, known as both MRTF-B and MKL2. MKL2, short for megakaryoblastic leukemia 2, is a member of the myocardin family (Selvaraj and Prywes, 2003). This family contains the protein myocardin (MYOCD), the transcription factors A and B (MKL1 and MKL2), and MASTR (Swärd et al., 2016).

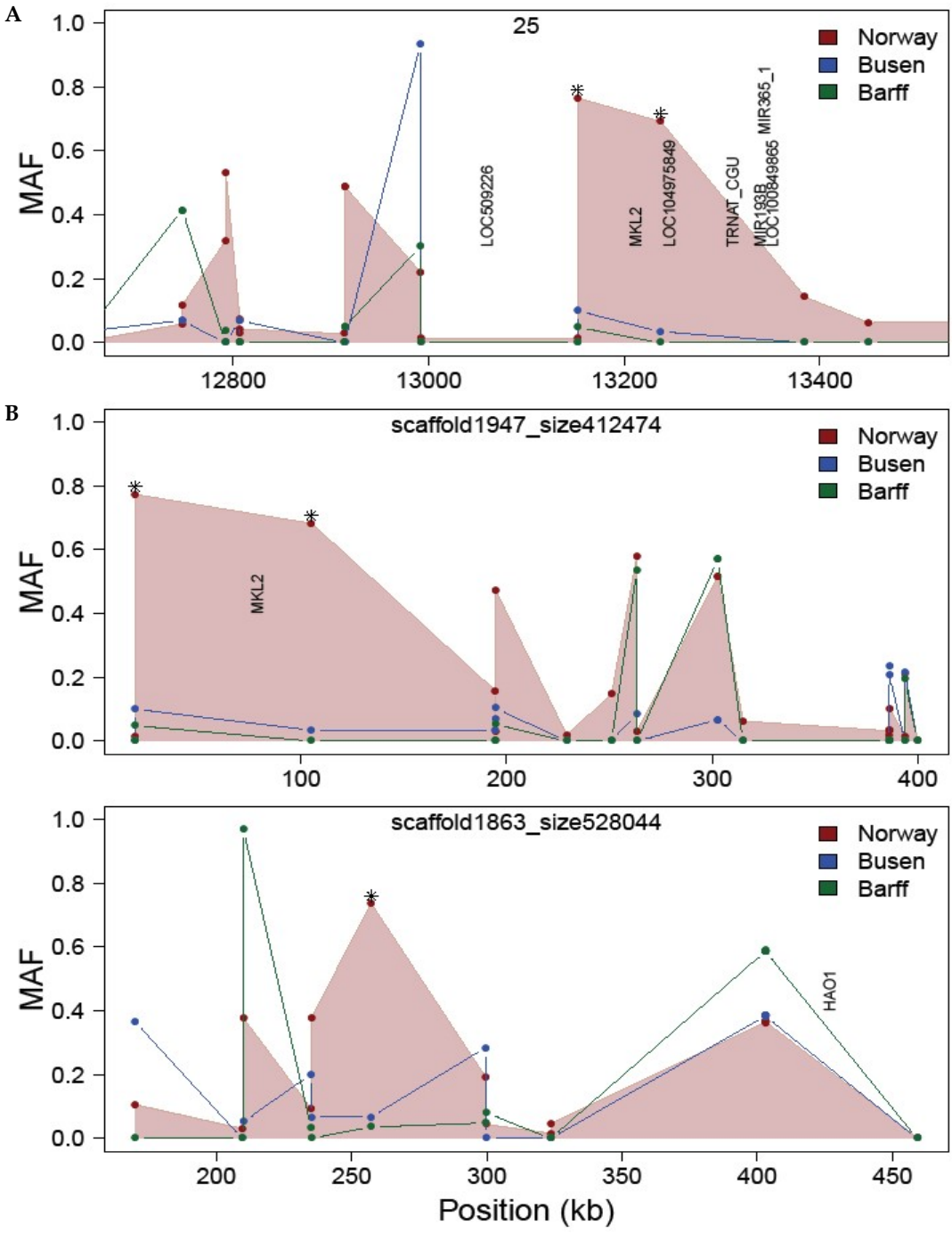


Fig.2.7. Genomic features close to outlier SNPs. *Genomic features within 200kB distance of the 3 outlier SNPs, according to alignment to both the cow genome.* ***A.*** *and the reindeer genome* ***B.****. Shading and lines show population specific minor allele frequencies of each SNP. Outlier SNPs are indicated with an asterix. Detected genomic features are uncharacterized loci LOC509226, LOC104975849, LOC100B49885, RNA sequences TRNAT_CGU, MIR193B, MIR365_1, and two protein coding genes: MKL2 and HAO1.*

MKL2 is a transcriptional coactivator of the serum response transcription factor (SRF). SRF controls the expression of muscle-specific genes, and is required for both striated and smooth muscle differentiation (Selvaraj and Prywes, 2003; Swärd et al., 2016). MKL2 is also implicated in E-cadherin-mediated cell-cell adhesion and signaling, which plays an essential role in development and maintenance of healthy epithelial tissues (Guo et al., 2014).

Discussion

In this study I capitalized on a semi-natural experiment to search for shared signals of selection in two sister populations which were simultaneously founded in geographically isolated but environmentally similar habitats. I additionally performed simulations to evaluate and provide additional support for my empirical findings. My overall aim was to gather empirical evidence that founder populations can start adapting directly following a founder event.

I screened the genomes of the study populations using four selection scans (i.e. Bayescan, GWDS, OutFlank and PCadapt) and two different approaches (i.e. pooled and pairwise approach) and found limited overlap in sets of loci marked as outliers. Most loci marked as outliers with the pooled approach were not marked as outliers with the pairwise/independent approach and vice versa. In addition, and as reported in previous studies (e.g. figure 2a in Andrew et al., 2018; figure 3a in Chen et al., 2018), most loci marked as outliers by one selection scan were not marked as outliers by other selection scans.

To better understand the observed inconsistencies, I ran simulations using a custom-built Wright-Fisher model simulator. This tool was specifically designed to simulate unlinked neutral and adaptive allele frequencies in founder and source populations following a founder event. I validated the model by comparing simulation results (i.e. proportion of retained alleles and fixation probability and time) with established equations from theoretical population genetics (Fig 2.4) based on the Wright-Fisher model.

My simulations provided estimates of the power and specificity of three R software packages for selection analysis (GWDS, PCadapt and OutFlank) in the context of pairwise source and founder population comparisons in the absence of gene flow. I evaluated the performance of each test for various combinations of

selection coefficients (s) and founder effective population sizes (Ne) (Fig 2.5) in recently diverged populations. The focus of this simulation study differs both in the methodology and aim from earlier simulation studies, which mainly evaluated the performance of selection scans under varying demographic models (De Mita et al., 2013; Lotterhos and Whitlock, 2014; Luu et al., 2017; Narum and Hess, 2011). The outcome of my simulations are only informative for the particular case of heterogeneous selection on standing variation in source-founder populations in the absence of gene flow, and caution should be exercised when extrapolating the results to other demographic scenarios.

My simulations indicated strong dependency of the performance of all three selection scans on both factors (s and Ne), with poor power resulting from low Ne and/or low s, exarcebated by the sampling effect. For Ne <= 50, the majority of positively selected loci were not detected by any selection test, unless the selection coefficient was very high (s >= 0.15). My simulations suggest relatively low power in small founder populations for the software OutFLANK. Zero power for OutFLANK under certain scenarios has been reported previously (e.g. figure 5 in Bernatchez et al., 2016; Luu et al., 2017). For founder populations with Ne ≥ 50, my simulations confirmed the claim of OutFLANK developer's that OutFLANK has high specificity without greatly compromising power (Lotterhos and Whitlock, 2015).

Visual examination of He-F_{ST} plots reveal that the low power of selection scans in small isolated founder populations does not reflect flawed test designs, but rather the confounding effects of genetic drift both during (bottleneck sampling) and after the bottleneck. Genetic drift can make selected loci indistinguisable from neutral loci in two ways: by affecting the detectability of selected loci directly and indirectly. Drift works directly on the selected loci itself, and can moderate or even counteract selective driven allele frequency change. In addition drift affects neutral loci and as such the backdrop of neutral variation from which selected loci need to stand out in order to be detected by selection scans (Lotterhos and Whitlock, 2015).

The indirect obscuring mechanism is especially relevant under two conditions: low Ne, and no gene flow. In small isolated populations the time window in which positively selected loci can stand out from the backdrop of neutral variation (i.e. approach and reach fixation before neutral alleles do so) is limited or near absent (Fig. A2.12). In big populations, in contast, neutral alleles take a long time to

reach fixation, which provides a wide time window for adaptive loci to stand out. In communicating populations (i.e. in the presence of gene flow), the allele frequencies in populations are correlated, and F_{ST} values do not converge to 1, resulting in an infinite time window in which heterogeneous selection can make adaptive loci stand out (figure 3a in Beaumont and Nichols, 1996).

The direct obscuring mechanism operates regardless of population size and gene flow, and can make selected loci indistinguisable from neutral loci despite the potential to stand out (Fig. A2.12). This effect can be either temporary or, if caused by loss of the adaptive allele, permanent, the latter possibility especially likely in small populations (Fig A2.12). Negative results from outlier tests could reflect the absence of selection, but also the influence of direct and indirect obscuring mechanisms, and should therefore not be overinterpreted (Lotterhos and Whitlock, 2015).

The direct and indirect obscuring mechanisms explain the presence of false negatives. It is less clear how the effect of drift causes false positives. I found that the majority of the loci marked as outliers by the selection tests for the empirical datasets were indistinguisable from neutral loci with regard to locus specific Weir & Cockerham He-F_{ST} scores (Fig 2.3D). My simulations indicate that false negatives are predominantly found on or in proximity to the lower left boundary of the He-F_{ST} spectrum (Fig. A2.14), the boundary reflecting fixation or loss in either population. In contrast, the empirical outlier SNPs are more widely scattered across the He-F_{ST} spectrum (Fig 2.3D).

It could be argued that a positively selected locus can have an ordinary He-Fst score and yet stand out in other respects. The finding that outlier He-F_{ST} scores cluster by selection scan (Fig 2.3D), suggests the probability of a locus being marked as outlier depends partly on the selection test used (i.e. Bayescan, GWDS, PCadapt or OutFlank). This might be suggestive of flawed test designs, but it might also indicate complementarity among selection scans. As evidenced by the existence of many different types of selection scans (Oleksyk et al., 2010), selected loci can exhibit various sorts of signals of selection. If these signals are uncorrelated, selection scans which query different signals will output different (i.e. complementary) sets of outliers.

My simulations however indicate that the inconsistencies between outputs of different selection tests do not result from complementarity, but are more generally indicative of type I errors. My simulations indicated that loci marked as outliers by only one test are predominantly false positives. In contrast, loci marked as outliers by multiple tests are predominantly true loci under selection. In a simulation of founder populations with a demographic scenario mirroring the demographic history of South Georgia study populations, and given a sample size of 30 individuals per population, most loci identified as outlier by only one test were false positives, whereas most loci identified by at least two selection scans were true adaptive loci (Fig 2.6C). More specifically, all loci identified as outlier by all selection scans were true adaptive loci (Fig 2.6C).

My simulations also provided insights into the observed inconsistencies between the pooled and pairwise approach, and suggested that these inconsistencies are more commonly indicative of false negatives than of false positives. In simulated populations with demographic histories similar to that of my study populations, only a minority of simulated adaptive loci were detected by both approaches (Fig 2.6A). This implies that a locus does not have to be detected by both approaches in order to be considered a true outlier.

I reasoned a priori that since the South Georgia reindeer populations might have underwent parallel evolution, they potentially shared genetic fingerprints of selection, which would increase the ability to differentiate true loci under selection from false positives. I realize that focusing on shared signals comes at the expense of overlooking private signals. Given the substantial loss of genetic variation in both populations (i.e. less than 65% retained variation, Fig 2.2E), a minority of adaptive alleles (i.e.: $0.65^2 = 0.4225$) is expected to be present in both populations. The implication is that most selective events are expected to be private events, occurring in either population but not both. My analyses however revealed a lower specificity of selection scans when applying the pairwise approach compared to applying the pooled approach (Fig 2.6A), offering less confidence in differentiating between false positives and true unshared loci under selection. As the main aim of my study was to provide compelling empirical evidence for selective events in founder populations, I therefore focused on results obtained with the pooled approach. In

other words: I directed my attention towards firmly established outliers, at the expense of overlooking less well-established outliers.

The insights gained from my simulations assist the interpretation of my empirical findings. The pooled approach resulted in three loci which were marked as outliers by two or more selection scans. Two of these loci were detected by both Bayescan and GWDS, whereas the third was identified by Bayescan, GWDS and PCadapt (Fig 2.3C). My simulations indicated that false positives are uncommon among loci detected by two and especially by three selection scans (i.e. Bayescan, GWDS and PCadapt), and therefore imply these three loci are most likely true loci under selection (Fig 2.6E-F). My simulations furthermore show that this conclusion is not contradicted by the fact that these loci were not detected by either of the pairwise comparisons (i.e. Busen-Norway and Barff-Norway, Fig 2.6A).

A potential confounding factor which cannot be assessed through simulations is the effect of genotyping errors. However, the relative positioning of the outlier loci argues against explanations involving genotyping errors, at least for the two adjacent SNPs. These two adjacent SNPs share a congruent signal of selection (Fig. 2.7) despite being located on different sequencing reads. The improbability of any pair of unrelated outlier SNPs being adjacent by chance, given the small proportion of outlier SNPs (3 out of 67.718 SNPs in total), greatly diminishes the chance that their unusual high F_{ST} values result from sequencing errors.

I observed that one of the reads containing an adjacent outlier SNP, contained next to the outlier SNP a neutral SNP. This neutral SNP, 34 bp distant from the outlier SNP, had population specific MAFs of 0, 0, and 0.01 for respectively Busen, Barff and Norway (source population) (Fig 2.7). The minor allele was possibly linked to the adaptive allelle, as the only copy in the source population occurred in an individual which was heterozygous for both the neutral SNP and the outlier SNP. But even if it was linked, the low number of copies within the source population makes it likely that this allele was lost in both founder populations, either due to the bottleneck or due to genetic drift in subsequent generations, before it could rise in frequency due to linkage. Hence, the presence of this neutral locus in the close vicinity of an outlier SNP, does not question the integrity of the outlier SNP.

Based on the analyses of my empirical and simulated datasets, I conclude that my study provides compelling empirical evidence that founder populations can adapt to their novel environment within ecological time scales. Theory predicts that founder populations have constrained adaptive capacities as a consequence of the founder bottleneck, which causes both a reduction of genetic variability (i.e. reduction of adaptive potential) (Willi et al., 2006) and a temporal increase of the magnitude of genetic drift. My simulations indeed indicate severe loss of genetic variation within the South Georgia founder populations, which makes that only a minority of potential adaptive alleles – less than 0.4225, as explained above – can be expected to have been retained both South Georgia founder populations (instead of in one population only). Depending on the level and nature of genetic variation within the source population, this can however still provide plenty of potential for parallel adaptive evolution.

With regard to the second adaptive constraint of founder populations – increased magnitude of genetic drift due to small population size – my simulations indicate that even in the face of strong genetic drift, selection of sufficient strength (e.g.: s = 0.1) can drive a proportion of adaptive alleles to fixation within a relatively short timeframe (i.e. 20 generations) (Fig A2.12; A2.14). In fact, it can even be argued that under certain conditions adaptive alleles have relatively high fixation probabilitys in founder populations. Imagine for example an adaptive allele (s = 0.01) which is represented by 10 copies in a diploid population of 1000 individuals, and which after a founder event is represented by 2 copies in a population of 5 individuals. According to Kimura's fixation probability function – i.e. u(p) = (1-exp(-4*N*s*p))/(1-exp(-4*N*s) (Kimura, 1962)– the fixation probability of this allele went up from 0.18 in the source population to 0.21 in the founder population. The reason of this increase is biased sampling: even though only 2 out of 10 adaptive allele copies were retained in the founder, the frequency of the allele went up from 0.5% to 20%. Because the fixation probabilities of deleterious alleles are especially likely to go up during a founder event (if retained in the founder population), purging of slightly deleterious alleles might represents a big challenge for bottlenecked (founder) populations (Feng et al., 2019).

If the identified outlier region(s) are indeed true loci under positive selection, the next question is: what were the associated phenotypic traits under selection?

Insular populations, such as the South Georgia reindeer, exhibit evolutionary trends in both morphological and behavioural traits (Losos and Ricklefs, 2009). One of these trends, the island rule or Foster's rule, involves changes in body size and predicts dwarfing of big species and gigantism of small species (Foster, 1964; Lomolino et al., 2013; Rozzi and Lomolino, 2017). Case studies of both extinct (e.g.: Lister, 1989) and extant species (e.g: Gray et al., 2015) illustrate that these changes can occur rapidly.

Cervidae are among the taxonomic groups which are particularly susceptable for insular dwarfing (Lomolino et al., 2013). Insular populations of reindeer are often characterized by reduced leg length, most extremely the Svalbard reindeer (Klein et al., 1987). Mainland populations adhere to Allen's rule by exhibiting a latitudinal gradient of decreasing leg length from south to north (Klein et al., 1987). These mainland and insular trends are thought to represent a trade-off between costs and benefits associated with long legs. Long legs provide increased locomotion efficiency and speed, which aids migration and predation avoidance, especially in deep snow cover. Long legs are however costly to build and to maintain, and complicate thermoregulation and foraging at ground level (Klein et al., 1987). There is however no evidence for decreased leg lengths in the South Georgia populations (Leader-Williams, 1988).

Rather than being associated with insularity, it is also possible that the trait under selection in the South Georgia populations were associated with factors specific for the South Georgia habitat. Environmental differences between South Georgia and the habitat of the Norwegian source population included a higher salinity (sea spray and greater proportion of marine grasses), the absence of predators, a milder climate (although with more heavy winds, (Leader-Williams, 1988, page 36), and dietary changes due to vegetation differences. According to investigations by Leader-Williams (1988), this latter category might have led to increased mortality rates among the South Georgia reindeer.

South Georgia reindeer mortalities followed patterns typical for deer, with females dying mostly in late winter and males mostly dying in early winter, after the rut (Leader-Williams, 1988). There were, however, two unusual mortality factors, not commonly observed in for insular populations, nor in the Norwegian source reindeer population. One unusual mortality factor consisted of falls over cliffs

(Leader-Williams, 1988). This occurred among all age classes, but especially in calves (Leader-Williams, 1988). The second unusual mortality factor was from dental disease.

Both South Georgia reindeer populations were affected by dental and mandibular abnormalities (Leader-Williams, 1988). Symptoms varied from missing to split or broken mandibular premolars and molars, regularly accompanied by mandibular swellings (Leader-Williams, 1982). These mandibular swellings affected 9-19 percent of all individuals within both populations (Leader-Williams, 1982, table 1). As mandibular swellings are likely to reduce the efficiency of chewing and therefore energy uptake, they could affect survability. Indeed, significant differences in both body condition and mortality rates were observed between affected and unaffected individuals (Leader-Williams, 1982).

Leader-Williams (Leader-Williams, 1982, table 3) found that 22.9% of over 100 examined carcassed were affected, whereas based on the prevalence in either population a percentage of 15.1% was expected. Field observation also suggest that affected individuals coupled their higher mortality rates with lower fecundity (Leader-Williams, 1988, page 177).

Both radiographic and chemical analyses show severe osteoporosis of mandibles, increasing with age and being more pronounced in individuals with mandibular swellings (Leader-Williams, 1988, page 174). Leader-Williams (1988) hypothesized a scenario in which a combination of overpopulation and limited availability of nutrient rich vegetation led to mineral deficiencies in the South Georgia reindeer. This caused osteoporosis in mandibulars, and increased suspectibility for tooth damage and tooth loss (Darcey et al., 2013). Tooth damage, which in turn predisposed affected individuals to swellings, may have been caused by increased suspectibility for infections by micro-organisms (Leader-Williams, 1988, page 175). I hypothesize that the South Georgia reindeer possessed heritable variation in susceptibility for mandibular osteoporis and tooth damage, resulting from the presence of a polymorphism within MKL2 itself or within a cis-regulatory element. Although I underscribe that mineral deficiencies in the newly colonized environment could explain the sudden manifestation of a previously unseen condition, I also remark that genomic stress resulting from bottlenecks can impact morphology as well (Lovatt and Hoelzel, 2011).

The exact mechanism through with an MKL2 allele could have counteracted mandibular osteoporosis and tooth damage despite mineral deficiencies, is unknown, and hypothesized mechanisms are speculative by nature. However, I propose that MKL2 variants might infer increased resistance to tooth disease by acting upon E-adherin (Guo et al., 2014). E-cadherin-mediated cell-cell adhesion and signaling plays an essential role in development and maintenance of healthy epithelial tissues (Guo et al., 2014). Teeth have a mesenchymal as well as an epithelial component, and E-cadherin is thought to regulate odontogenesis (Heymann et al., 2002; Li et al., 2012).

The proposed scenario corresponds to the type of substitution events envisioned by Haldane (1957), which considered a population which 'due to deteriorating circumstances, finds a previously satisfactory gene inadequate so that it comes to be replaced by a previously neutral or undesirable allele which had remained rare' (Brues, 1964). In this scenario, the genetic load (i.e. the difference between reference optimal fitness and actual fitness) experienced by the population does not result from mutation pressure, but instead from external factors, namely environmental change. The proposed scenario does therefore not imply a genetic load in the reindeer source population, as the genetic polymorphism could have been neutral prior to the colonization of the new environment.

As pointed out by Haldane (1957), fitness reduction due to environmental change is accompanied by a reduction in population size. The extent and duration of the population size decrease depends on the presence of potentially adaptive standing genetic variation and/or waiting time to arrival of new beneficial mutations. If a population contains a genetic variant which, given the new environmental circumstances, has a higher fitness than the originally dominant allele, fixation of this new allele would be accompanied by a population size increase. The net outcome of these dynamics – on the one hand the lowering of the fitness of 'wildtype' individuals which leads to a population size decrease, and on the other hand the fitness gain of 'mutant' individuals causing a population size increase – depends on the magnitude of change of both selection coefficients. If the fitness gain of the mutant phenotype is higher than the fitness loss of the wildtype phenotype, the population will eventually increase in size. But whereas reproduction and population growth of reindeer populations occur over a time-scale of years,

mortality associated with environmental change can occur instantly. It is therefore likely that even if the net long-term outcome would be a population size increase, the initial response would be a population size decrease.

In the case of founder populations, such as the South Georgia reindeer, the population dynamics are affected by the bottleneck event. It has been argued that underpopulation can lead to a relaxed selection regime, and that during a population expansion following a bottleneck event (the so called 'flush' – Carson, 1968) a low-fitness alleles can rise in frequency (Carson, 1968). However, although during underpopulation (in which a population is below it's carrying capacity) the absolute fitness of negatively selected individuals can indeed exceed 1, the relative fitness of these individuals will be below 1. Individuals carrying the deleterious allele will multiply more slowly than individuals carrying the advantageous allele, and therefore the frequency of the deleterious allele will decrease (as long as not counteracted by drift). Eventually, when approaching the population carrying capacity, the absolute fitness of negatively selected individuals will drop below 1, and their numbers will decrease. Possibly, the results presented in this study provide empirical evidence for such a substitution process during a population size expansion.

Conclusions

My simulations show that for sister founder populations subjected to similar environmental conditions, positively selected loci are more confidently detected by the newly developed selection scan GWDS compared to the widely used selection scans Bayescan, OutFlank or PCadapt. I detected 3 SNPs - 2 of which were adjacent to each other, and all three marked as outlier by two or more selection scans – with fingerprints of positive selection in two heavily bottlenecked deer founder populations of less than 10^2 years old. Wright-Fisher model simulations provide further support that these 3 outlier SNPs are true loci under selection. The genetic signals of selection could correspond to differential survival rates among individuals with and without mandibular swellings resulting from dental disease. This study therefore provides empirical evidence that despite their adaptive constraints founder populations can start adapting to their novel environment directly following a founder event.

Chapter 3

Demographic and evolutionary history of the native UK roe deer (*C. capreolus*) population inferred from ddRADSEQ SNP data

Abstract

The British mammalian fauna is similar to that of north western mainland Europe, both in terms of species composition and in terms of species characteristics. The similarity in species composition can be traced back to the existence of a Holocene land bridge, Doggerland, which allowed recolonisation of the British Isles following the Younger Dryas. The apparant similarity in species characteristics might reflect absence of diversifying selection, but adaptive traits are often obscure. In this study I harnassed the ddRADseq protocol to generate SNP datasets of European roe deer populations occuring on either side of the North Sea in order to analyse the extent, and causes, of the genetic differentiation of the native UK roe deer population from the mainland population. My analyses indicate that the effective population size of the native UK roe deer population has numbered a few thousand individuals throughout its separate history, resulting in moderate levels of genetic drift which have led to moderate loss of standing genetic variation. Selection scans revealed the existence of two adjacent outlier SNPs (out of over 50K SNPs in total) which possibly experienced diversifying selection. Neither genetic drift nor diversifying selection has however been sufficient to cause fixed differences between the native UK and mainland roe deer populations.

Related peer-reviewed publication:

De Jong, M.J., Li., Z., Qin, Y., Quemere, E., Baker, K., Wang, W. 2020. *Demography and adaptation promoting evolutionary transitions in a mammalian genus that diversified during the Pleistocene*, Molecular Ecology

Author contributions:

ARH conceived the study and MdJ & ARH wrote the paper. MdJ undertook data and lab analyses. EQ provided a subset of the RADseq data. ZL, YQ and WW generated the C. pygargus genome assembly and annotation.

Introduction

Continental islands that are presently separated from the adjacent mainland by seaways shallower than 120m, were connected to the mainland during the Last Glacial Maximum (LGM; Burridge et al., 2013). As the timing of sea level changes is generally well known (Lambeck and Chappell, 2001), the maximum age of insular populations on continental islands can be precisely estimated, which facilitates inferences about the evolutionary history of these populations and about evolutionary processes in general (Comes et al., 2008; Lister, 2004; Velo-Antón et al., 2012).

Continental islands which formed after the LGM contain few endemic species and exceptions often represent relict endemics (e.g. Brown, 2006), indicating that a typical speciation duration exceeds $2 \cdot 10^4$ y (Lister, 2004). Endemic subspecies, in contrast, are common on continental islands, illustrating that subspecies can form within relatively short time spans. An abundance of dwarfed and giant (sub)species on continental islands showcase selection driven ecological divergence between mainland and insular population (Lomolino et al., 2013). Although continental islands have experienced multiple cycles of sea level changes throughout the Pleistocene (Burridge et al., 2013), in many instances it can be inferred that present day variation became established after the LGM. Well studied cases include the Svalbard reindeer (Klein et al., 1987), the Tasmanian emu (Thomson et al., 2018), the Channel island fox (Funk et al., 2016; Hofman et al., 2015; Robinson et al., 2016), the Cozumel pygmy raccoon and dwarf coati (McFadden et al., 2008), and Australian tiger snake (Keogh et al., 2005).

Body size ranks amongst the most easily identifiable species traits. In theory these observed body size differences might represent the top of the adaptation iceberg, and other more obscure adaptive traits might remain to be discovered. Dense SNP catalogues allow to screen genome wide genetic variation and to search for adaptive driven differences between insular and mainland populations (Haasl and Payseur, 2016).

The British Isles are landbridge islands which were cut off from continental Europe after the LGM. Unlike Ireland, which became an island around 15 kya (Montgomery et al., 2014), Great Britain was connected to the mainland until relatively recent. This connection comprised a landbridge known as Doggerland,

which is nowadays submerged under the southern North Sea and which was flooded approximately 6-7 kya (Coles, 1998; Sturt et al., 2013). Doggerland facilitated the recolonization of Great Britain by temperate species after the Younger Dryas (i.e. < 11.7 kya; (Coles, 1998). As a result, the faunal composition of Great Britain is very similar to the faunal composition of north western Europe (Montgomery et al., 2014; Stuart, 1995).

The faunal similarity on either side of the North Sea includes similarity in species appearance. This phenotypic similarity of native British populations to their mainland counterparts could reflect the absence of diversifying selection. Adaptive differences can however be subtle and obscure, and therefore the influence of diversifying selection can not be ruled out based on apparent absence of phenotypic and niche differentiation alone. Scrutinious examination of genomic wide genetic differentiation has previously identified putative adaptive traits within a British population which otherwise might have remained undetected (Bosse et al., 2017).

In this study I aimed to obtain more insight into the evolutionary history of native British populations by focussing on one of the biggest extant native British mammals: the European roe deer (*Capreolus capreolus*). This species has been present in Europe for at least 600 ky (Andersen et al., 1998), of which in Britain during interglacials (Stuart, 1995). As typically observed for temperate Pleistocene mammals, the roe deer fossil record provides evidence for range contractions to refugia during glacials and subsequent range expansions during interglacials (Sommer and Zachos, 2009; Sommer et al., 2009). Mitochondrial DNA and microsatellite DNA studies have indicated that during the LGM at least four such refugia were present and that a refugium in central Europe served as the main base for recolonization of north western Europe and Great Britain (Baker and Hoelzel, 2014; Hewitt, 1999; Randi et al., 2004). The fossil record furthermore suggests that roe deer were absent from Doggerland and Great Britain during the Boling-Allerod interstadial and the Younger Dryas and first appeared during the early Holocene (Van Kolfschoten and Laban, 1995), perhaps dictated by the spread of broadleaved forests (Baker and Hoelzel, 2014; Petit et al., 2003).

In this study I applied the double digest restriction-site associated (ddRAD) sequencing protocol to generate genome wide SNP datasets of four roe deer populations distributed on either side of the North Sea. My aim was to obtain better

insight in the demographic and evolutionary history of the native UK roe population. I was particularly interested in two questions: 1.) what was the effective population size of the roe deer population which colonized Great Britain?; and 2.) has the genome wide genetic divergence of the British and mainland roe deer populations been affected by natural selection?

Methods

Sample collection. I collected tissue samples of roe deer from four sampling localities of comparable size, of which two were located in western mainland Europe and two in the United Kingdom. I chose a sampling location in Wurttemberg, Germany, to represent the central European roe deer lineage from which the native UK roe deer population derived. I chose a sampling location in Ayrshire, Scotland, to represent the native UK population. Roe deer were hunted to local extinction in England during medieval times and have recolonized England since, both naturally (through migration out of Scotland) and artificially (through anthropogenic reintroductions, stocked from mainland Europe) (Baker and Hoelzel, 2014). The native UK roe deer population is therefore better represented by a Scottish population than by a English population.

The other two sampling locations were included for contrast. I included a sampling location in southern France, Aurignac, which allowed us to compare the genetic differentiation of the UK roe population to the genetic differentiation between mainland populations. Secondly, I also collected samples from a roe deer population which split from the Wurttemberg population recently and which was affected by a severe population bottleneck. This human-made population was founded around 1880 with the translocation of 10 individuals from Wurttemberg to East Anglia, England (Baker and Hoelzel, 2014). I included this population to gain insights into the genetic differentiation of a heavily bottlenecked population, providing a contrast to the genetic differentiation of native the UK population, and allowing to assess the impact of a well documented bottleneck on genome wide variation.

Samples were collected during culls, or reused from earlier studies (Baker and Hoelzel, 2014; Gervais et al., 2019). No animals were killed specifically for either of these studies, and all animals were killed by certified stalkers.

DNA Extraction and Library Construction. Libraries were constructed following the ddRADseq protocol and paired-end sequenced using an Illumina HiSeq_2500 (version 4 chemistry) machine. For the Ayrshire, East Anglia and Wurttemberg (AEW) dataset I used a 6 bp cutter (*HindIII*: AAGCTT) and a 4 bp cutter (*MspI*: CCGG), with a fragment size selection window of 250 bp width (including all fragments with a length of 275 to 525 bp, excluding the adapters). Based on in silico simulations with the R package SimRAD, I expected to extract 120,000 loci with an average read depth of 30. By multiplying this expected number of loci against their average length (250 bp), a conservative estimate for nucleotide diversity ($\theta = 1/2000$), and an approximation for the harmonic number of Watterson's estimator, I estimated that this size selection window would yield at maximum ~50,000 SNPs with MAF > 0.05. The actual size selection was executed with a Sage Science PippinPrep machine. The Phusion High-Fidelity kit was used for a 13 cycle PCR (denaturation step: 62°C for 20sec; annealing step: 72°C for 45 sec; extension step: 72°C for 5 min).

The Aurignac dataset, which was created independently for another study (Gervais et al., 2019), was generated with the same frequent 4 bp cutter (*MspI*) but with a different 6bp cutting enzyme (*EcoR1*), and with a fragment size selection window of 60 bp width (including all fragments with a length of 270 to 330 bp, excluding the adapters).

SNP calling and filtering. Reads were demultipled and trimmed to 110 bp (or 117bp in the case of the Aurignac dataset (Fig S.3.1) using the software STACKS version 1.35 (Table A3.1). Unpaired reads were discarded. Paired reads were aligned against both the newly generated *Capreolus pygargus* genome (see Chapter 4 of this thesis) as well as the *Cervus elaphus* genome (Bana et al., 2018) using the software Bowtie version 2.2.5. I chose the red deer genome as a second reference genome because red deer is the species closest to *Capreolus* with a genome assembly up to chromosome level. Samtools version 1.3.3 was used to filter out reads which aligned to more than one location in the genome, which aligned disconcordantly; and those with a mapping quality below 20.

SNPs were called using the STACKS refmap pipeline with default settings. Loci for which at least 30 percent of all individuals had a read depth below 8 were

removed. I accepted multiple SNPs per read (i.e. I did not set the –write-single-SNPs flag when running the 'populations'-command), as I opted to 'thin' the dataset downstream.

PGDSpider and PLINK v1.90 were used to convert the output from genepop format to a genlight object, implemented in the R package Adegenet, and the tool 'depth' of vcftools (Danecek et al., 2011) was used to calculated read depth among samples and among SNPs.

I filtered the SNP datasets on proportion of missing data, heterozygosity excess, minor allele count and on read depth. To be more precise, I excluded samples with more than 25 percent missing data, and subsequently sites with more than 10 percent missing data, sites with unusual high deviation from Hardy Weinberg expectations (Fig. A3.3), sites with only one copy of the minor allele. and all sites belonging to the 1% class of loci with the highest read depths (Fig. A3.4). I also filtered out a small number of SNPs which mapped to the same location of the reference genome, even though they belonged to different STACKS loci. I additionally thinned the dataset by selecting at maximum 1 SNP per 500 bp window.

I extracted the intersect of the two SNPs datasets (i.e. the dataset containing Ayrshire, East Anglia and Wurttemberg samples vs dataset containing Aurignac samples) based on the locations of the SNPs in the reference genome (*Capreolus pygargus* genome, see Chapter 4 of this thesis).

For the selection analyses, I used the filtered, non thinned roe deer aligned AEW dataset. For genetic diversity analyses, I used the filtered and thinned datasets of both the AEW and Aurignac datasets. For genome wide genetic diversity analyses, I used the filtered, non thinned red deer aligned AEW dataset. For population structure analyses, I used the filtered and thinned datasets of the AEW dataset and the intersect dataset.

Population genetic analyses. Nei's genetic distance, admixture and structure analyses, as well as site frequency spectra (SFS) and genotype network construction, were executed in R, using either in-house-built functions or functions implemented in the adegenet, Ape, StaMPP, LEA, Poppr (Kamvar et al., 2014) and PEGAS packages. The package adegenet was used for data management and DAPC analyses, the package StaMPP for the calculation of Nei's genetic distance, the package LEA for

admixture analyses, the package Poppr for the calculation of Hamming's genetic distance, and the package PEGAS for the construction of genotype networks.

For DAPC analyses, executed using adegenet, I set the number of PCs to a third of the number of individuals – thereby ignoring the a-value, which suggested to retain 1 PC only –, the number of clusters to the number of populations, and the number of discriminant functions to 3. For admixture analyses, executed using LEA, I set K (number of clusters) to 2-6, alpha to 10, tolerance to 0.00001, and number of iterations to 200.

Contemporary gene flow was estimated using BayesAss3-SNPs. The number of iterations was set to 1,000,000, burn-in to 100,000 and delta values to 0.1. Relatedness between samples was calculated using plink version 1.90b3.38.

Population assignment test. I constructed and conducted within R a population assignment test using an approach similar to (but on same aspects different from) the approach described in Paetkau et al. (1995) and Pritchard et al. (2000). My approach calculates the probability that an individual belongs to a certain population given its observed genotype and given the minor allele frequencies within that population, as follows:

$Pr(popA|genotype) = Pr(genotype|popA)/(Pr(genotype|popA)+Pr(genotype|popB))$

$Pr(popB|genotype) = Pr(genotype|popB)/(Pr(genotype|popA)+Pr(genotype|popB))$

For example, given two populations (A and B) which have for a particular locus a minor allele frequency of respectively 10 percent and 1 percent, the probability that an individual which is homozygous for both major alleles belongs to either popA or popB is estimated as:

$Pr(popA|0) = Pr(0|popA)/(Pr(0|popA)+Pr(0|popB))=(0.9\cdot0.9)/(0.9\cdot0.9+0.99\cdot0.99)= 0.45$

$Pr(popB|0)= Pr(0|popB)/(Pr(0|popA)+Pr(0|popB))=(0.99\cdot0.99)/(0.9\cdot0.9+0.99\cdot0.99)= 0.55$

For k loci, I calculated the probabilities Pr(geno|popA) and Pr(geno|popB) by multiplying each locus specific probability (assuming they are independent) as:

$Pr(geno|popA) = Pr(locus\_1|popA)\cdot Pr(locus\_2|popA)\cdot\ldots\cdot Pr(locus\_k|popA)$

$Pr(geno|popB) = Pr(locus\_1|popB)\cdot Pr(locus\_2|popB)\cdot\ldots\cdot Pr(locus\_k|popB)$

Excluded from the calculations were snps for which one of either alleles were not represented in either of the populations. Those loci would make the probability converge to 0 or 1, and hence were omitted.

Calculation of theta and genome wide heterozygosity. I estimated genetic diversity within the study populations by calculating pairwise sequence dissimilarity, the proportion of differences between two haplotypes. This metric can be derived from almost any population genomics datasets, and can be used to estimate genetic diversity within single individuals (i.e. heterozygosity) and within populations (i.e. nucleotide diversity (π) (Nei and Li, 1979) as well as genetic divergence between populations and even between species (see for example Table S5.2 in Malinsky et al., 2018; Fig 1B in Prado-Martinez et al., 2013). The use of this metric therefore facilitates comparisons among genomic datasets of various nature (as also stressed in Funk et al., 2016).

I calculated pairwise sequence dissimilarity as the average number of differences between haplotypes (as derived from genotype information). If haplotypes represented the two haplotypes of one individual, the pairwise sequence dissimilarity was effectively heterozygosity. In this latter case, I first calculated 'He_seg', the proportion of heterozygous sites within an individual relative to all sites which were segregating within the population to which the individual belonged. Second, I calculated genomeHe using the formula: genomeHe = (He_seg·N_seg)/N_total, in which N_seg equals the number of segregating sites and N_total equals the combined length of all loci/stacks which passed the STACKS filter settings. As value for N_total I used the value provided by STACKS in the sumstats_summary.tsv file. The total number of sites is listed in the third column ('Variant sites') of the second part of this file, after the line '# All positions (variant and fixed)'. Nucleotide diversity was calculated similarly, by calculating the mean number of differences for all possible pairwise sequence comparisons.

Stairway plots. The demographic histories of the study populations were inferred using the Stairwayplot analysis (Liu and Fu, 2015). I set the generation time to 5 years (Nilsen et al., 2009), and the mutation rate per site per generation to $1.1 \cdot 10^{-8}$. This estimate is based on the assumption that the mutation rate per year equals

$2.2 \cdot 10^{-9}$ (Kumar and Subramanian, 2002) and on the assumption that the mutation rate per generation relates linearly to the mutation rate per year. The population specific folded site frequency spectrum (SFS) vectors were generated with a custom-built script which binned SNPs in classes based on their number of copies of the minor allele, and subsequently calculated the size of each bin.

Selection analyses. Selection analyses were carried out according to the approach described in Chapter 2 of this thesis. One difference with the analyses in Chapter 2 was that I excluded Bayescan, and included a selection scan which is implemented in the R package Fsthet (Flanagan and Jones, 2017). As in Chapter 2, I applied both the independent and the pooled approach. For the pooled approach I used two different variants. In the first variant, which I labelled the modern UK-mainland comparison, I grouped samples from the Ayrshire and East Anglia population together and compared them against the German population. In the second variant, which I labelled the native UK-mainland comparison, I compared the Ayrshire samples against a group of samples belonging to both the East Anglia and the German population. Since GWDS, OutFlank and PCadapt are interpopulation scans which require high numbers of SNPs shared across two (or more) populations, I necessarily excluded the Aurignac population from selection analyses.

Genes nearby outlier SNPs were detected using the software bedtools2 and using the annotation file of the *C. pygargus* reference genome (see Chapter 4 of this thesis).

For simulations of expected locus specific He-Fst distributions of the native UK versus mainland comparison, I followed the same procedure as described in Chapter 2. Based on results obtained with the Stairwayplot analyses, I set the Ne of the founder UK population to 5000 (with no founder bottleneck) and the Ne of the European mainland population to 10000. The TMCRA of both populations was

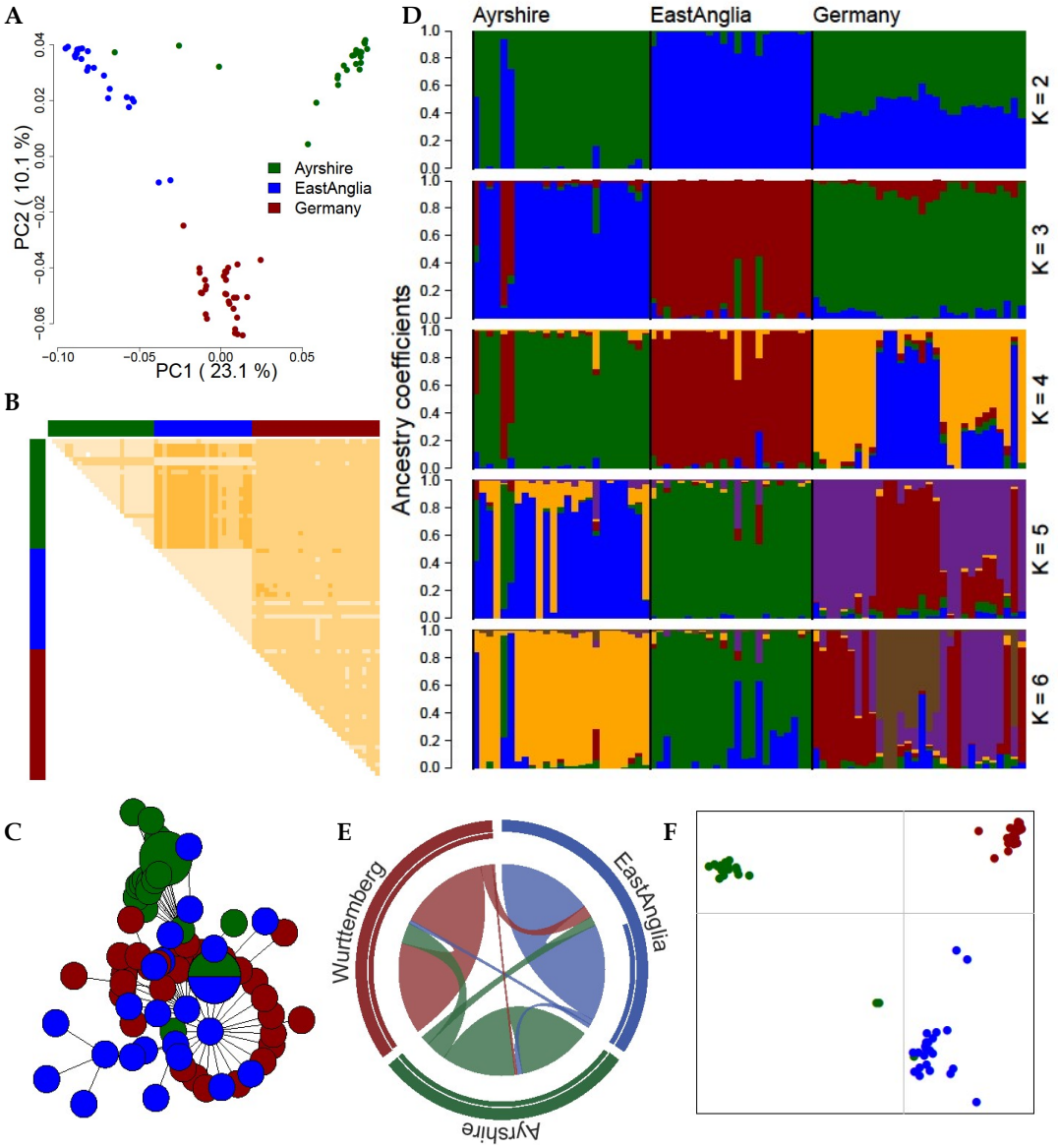


Fig. 3.1. Genetic clustering analyses of roe deer samples from three populations on either side of the North Sea. *Colour coding (except for D): blue = EastAnglia (introduced UK), green = Ayrshire (native UK), red = Wurttemberg (Germany), orange = Aurignac (France) .* ***A.*** *Principal coordinates analysis based on Nei's genetic distance (excluding Aurignac).* ***B.*** *Nei's genetic distance between samples (excluding Aurignac).* ***C****. Genotype network based on 286 snps among all four populations.* ***D.*** *Admixture analyses for 2 <= K <= 6, with random colour coding.* ***E.*** *Migration rates between the four populations, as inferred by Bayesass3-SNPs.* ***F.*** *DA1 vs DA2 of discriminant analysis of principal components, with nclusters set to 3.*

set to 1500 generations, which, assuming a generation time of 5 years, precedes the flooding of Doggerland, which is dated at ~6-7 kya (Coles, 1998; Sturt et al., 2013).

Selection analyses on a control dataset. For comparison, selection analyses (again according to the approach described in Chapter 2 of this thesis) were performed on a control dataset containing a locus which experienced a confirmed episode of positive selection. This control SNP dataset was a dataset of human samples and of SNPs of chromosome 2, obtained from the International Genome Sample Resource (IGSR, https://www.internationalgenome.org/data-portal). Chromosome 2 contains the gene responsible for lactose tolerance in north western European populations. Analyses were performed on a dataset of 30 GBR (Great-Britain), 30 FIN (Finland) and 30 TSI (Toscane) individuals, for a pooled comparison (GBR and FIN combined vs TSI) as well as two pairwise comparisons (FIN vs TSI, GBR vs TSI).

Results

SNP calling and filtering. The two sequencing lanes of the AEW dataset produced a combined number of 602.6 million single-end reads (Table A3.1). Almost 5.5 million reads had to be discarded due to either low quality or an ambiguous radtag, resulting in an average number of 6.8 million read pairs per sample (stdev: 5.2 million, min: 0.7 million, max: 23.4 million) (Table A3.1).

For the AEW dataset aligned to the roe deer genome, STACKS obtained 686,859 loci/stacks, of which 74,518 loci/stacks passed the filter settings ('sample/population constraints'), consisting of 8,196,980 sites, of which 52,364 (0.64%) sites were bi-allelic. The bi-allelic sites were concentrated on 34,250 loci/stacks. For the AEW dataset aligned to the red deer genome, STACKS obtained 434,524 loci, of which 44,934 loci passed the filter settings ('sample/population constraints'), consisting of 4,942,740 sites, of which 27,298 sites (0.55%) were biallelic, with on average (excluding SNPs aligned to Y-chromosome) 793 SNPs per chromosome (sd = 293).

For the Aurignac dataset STACKS obtained 259,987 loci, of which 50,975 loci passed the filters ('sample/population constraints'), consisting of 5,607,250 sites, of which 29,488 (0.53%) sites were bi-allelic. The bi-allelic sites were concentrated on 19,772 loci/stacks.

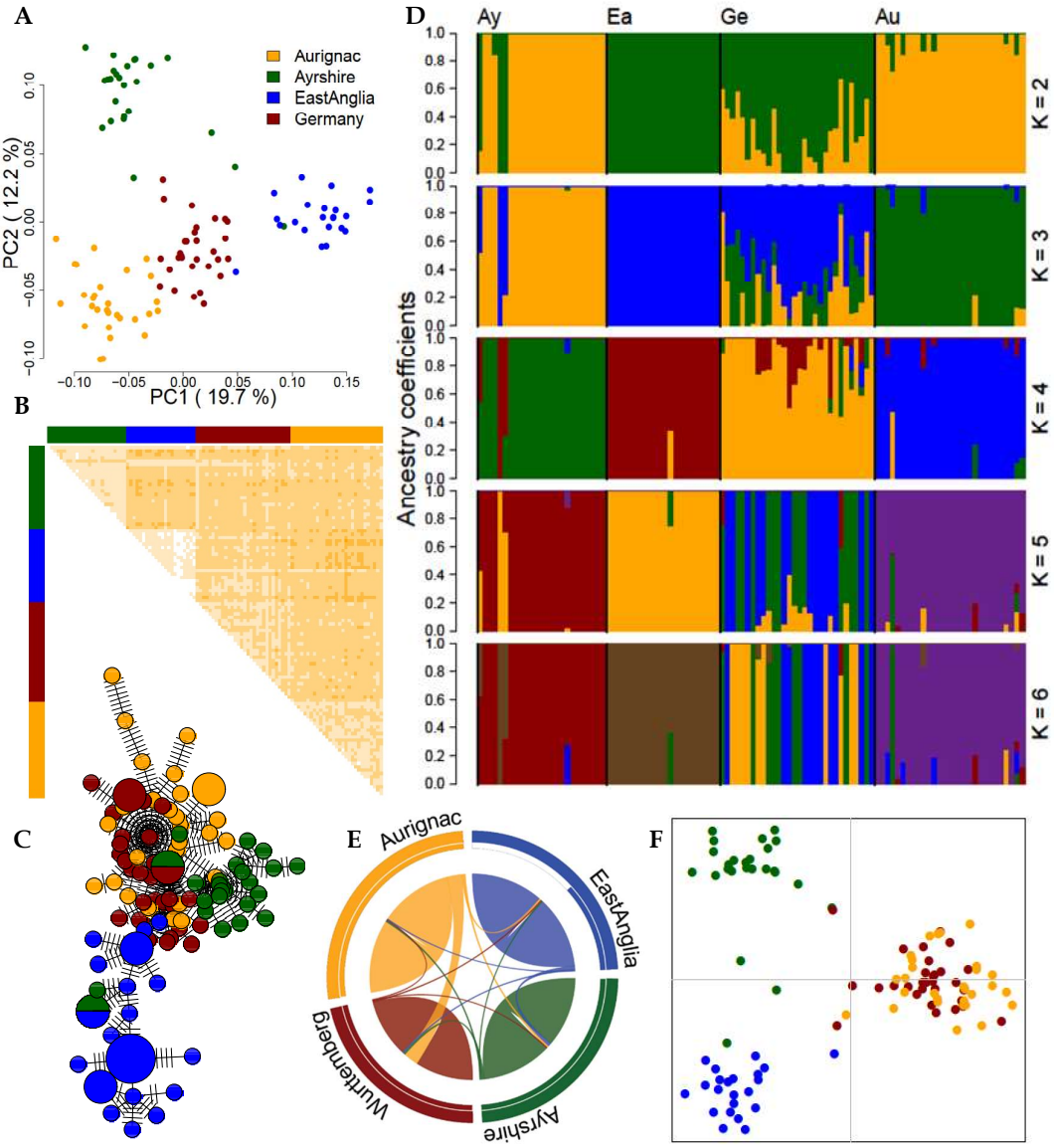


Fig. 3.2. Genetic clustering analyses of roe deer samples from four populations on either side of the North Sea. *Colour coding (except for D): blue = EastAnglia (introduced UK), green = Ayrshire (native UK), red = Wurttemberg (Germany), orange = Aurignac (France) .* ***A.*** *Principal coordinates analysis based on Nei's genetic distance (excluding Aurignac).* ***B.*** *Nei's genetic distance between samples (excluding Aurignac).* ***C****. Genotype network based on 286 snps among all four populations.* ***D.*** *Admixture analyses for 2 <= K <= 6, with random colour coding.* ***E.*** *Migration rates between the four populations, as inferred by Bayesass3-SNPs.* ***F.*** *DA1 vs DA2 of discriminant analysis of principal components, based on ~300 loci shared among all four populations, with nclusters set to 3.*

The coverage of the 52,364 SNPs of the AEW dataset followed a normal distribution, with a mean and median read depth of respectively 2727 (sd = 2314) and 2589, corresponding to a mean read depth per locus per individual of 29. The coverage of 29,488 SNPs of the Aurignac dataset also fit a normal distribution, with a mean and median read depth of respectively 2325 (sd = 1779) and 2169, corresponding to a mean read depth per SNP per individual of 78. The difference in read depths between both datasets reflected the differences in window size selection, and indicate that a size selection window of 250 width makes more efficient use of available sequencing resources than a size selection window of 60 bp width.

In line with expectations for paired-end sequencing data, the spacing between adjacent SNPs followed bimodal distributions. One modus represented SNPs occuring on read mates and another represented SNPs occuring on the same read (Fig. A3.2). The mean and median distances between adjacent SNPs per chromosome equalled 130.0±46.5 and 16.0±9.5 kbp respectively (±1 sd) (Table A3.2).

After filtering, I retained 107 samples, distributed over populations as follows: Ayrshire (Scotland) = 25, East-Anglia (England): 23, Wurttemberg (Germany) =30 and Aurignac (France) =29. For the AEW dataset I retained 31,459 SNPs after filtering and 15,697 SNPs after thinning (Table A3.3, Fig. A3.5-A3.7). For the Aurignac dataset I retained 19,992 SNPs after filtering and 10,732 SNPs after thinning (Table A3.5, Fig. A3.5-A3.7).

Overlap between datasets. The AEW and the Aurignac dataset were generated using the same frequent 4 bp cutter (i.e. *MspI*), but with different less frequent 6 bp cutter (i.e. *HindIII* and *EcoR1*). As the less frequent cutter determines which regions in the genome will be sequenced, overlap between both datasets was expected to be limited.

I found that the intersect of both datasets (i.e. 52,364 SNPs from AEW and 27,298 SNPs from Aurignac) consisted of 286 SNPs (≤1% of the SNP datasets). All 286 SNPs had the same allele pairs across both datasets, providing strong evidence that they were true shared SNPs. Out of the 286 SNPs, 258 SNPs were retrieved from the same position in the sequence read generated for either datasets, indicating that

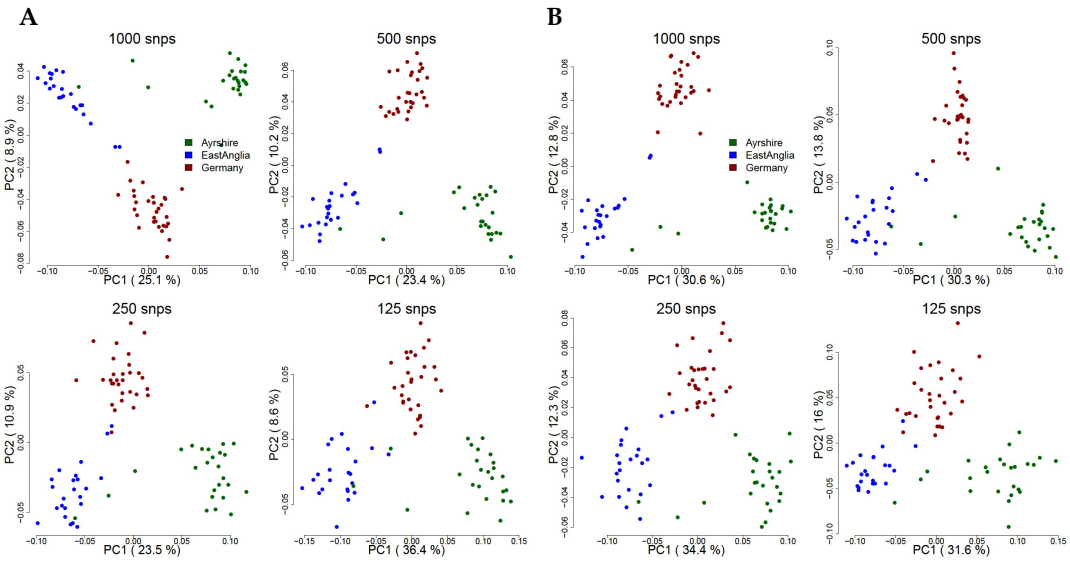


Fig. 3.3. Structure analyses on data subsets. *Principal Coordinate Analyses based on* ***A.*** *Hamming's genetic distance and* ***B.*** *Nei's genetic distance, for various sizes of random subsamples of SNPs.*

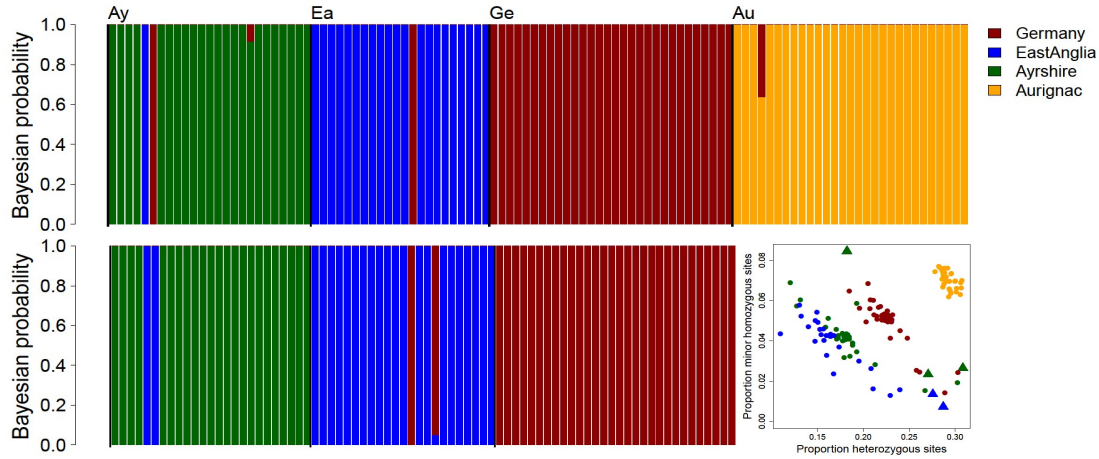


Fig. 3.4. Bayesian population assignment probabilities. *Colour coding: blue = EastAnglia (introduced UK), green = Ayrshire (native UK), red = Wurttemberg, orange = Aurignac (France). Above: results based on intersect dataset of 250 SNPs. Below: results based on full dataset of 15,697 SNPs. The scatterplot shows the proportion of heteryzogous and minor homozygous genotype calls per sample, both for samples which were assigned to the correct population (cirkels) and samples which were assigned to the incorrect population or were assigned to the correct population but with a probability below 1 (triangles).*

in either dataset the reads were sequenced starting from the frequent cutter (*MspI*) cut site. In 40 out of 286 SNPs the major allele in the AEW dataset was the minor allele in the Aurignac dataset. After filtering and thinning 250 SNPs were retained.

Structure analyses. Population structure analyses (i.e PCA, DAPC, genotype network, Nei's genetic distance and admixture analyses) indicated distinct population structuring, with each sampling locality clustering as a separate entity (Fig 3.1-Fig 3.2; Fig A3.8). This result was observed both for the AEW dataset (i.e. 15,697 SNPs, Fig 3.1) as for the intersect dataset (i.e. 250 SNPs, Fig 3.2). Reruns of PCA analyses on random subsamples of the AEW dataset confirmed that a relatively small number of biallelic SNPs (i.e >=125) suffices to infer main clusters within this particular dataset (Fig. 3.3). PCA analyses executed on variously sized subsample datasets of the AEW dataset confirm that a relatively small number of markers suffices to discern the correct population structure for the roe deer samples (Fig. 3.3).

A few samples did not cluster according to a priori expectations. Two East Anglia samples stood out by sharing similarities with Wurttemberg (Germany) samples, and three Ayrshire samples stood out by sharing similarities with East Anglia samples (Fig 3.1A, Fig 3.1F). The Bayesian population assignment test confirmed that based on the population allele frequencies and based on the genotype scores of the individuals, two East Anglia samples were more likely to belong to the Wurttemberg (Germany) population, and two Ayrshire samples were more likely to belong to the East Anglia population (Fig 3.4).

PCA and DAPC analyses indicated that the Ayrshire population is genetically more similar to the Wurttemberg (Germany) population (D = 0.061) than to the Aurignac population (D = 0.069), and that the Wurttemberg population is more similar to the Aurignac population (D = 0.048) than to East Anglia population (D = 0.069) (Fig 3.2, Fig 3.5). Around 30 percent of all SNPs were represented by private alleles in the Germany population, compared to 9% and 3% private alleles in respectively Ayrshire and EastAnglia (Fig 3.5B).

***Genetic diversity*.** The highest proportion of segregating sites was observed within the Wurttemberg (Germany) population, both before and after correcting for

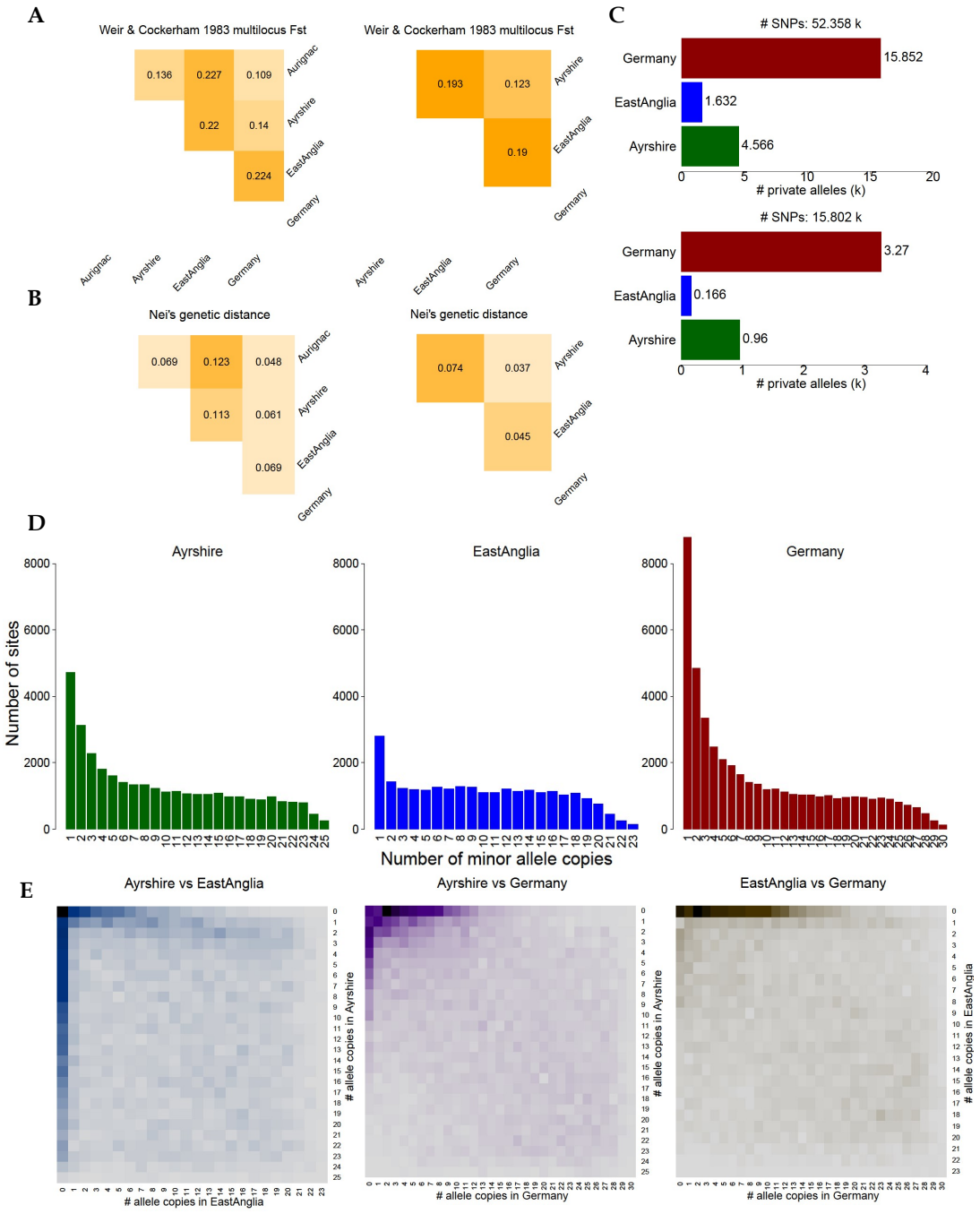


Fig. 3.5. Genetic distance and genetic diversity. A. *Multilocus Weir & Cockerham Fst-values for pairwise population comparisons in the AEW (right)and intersect (left) datasets.* ***B****. Nei's genetic distance for pairwise populations comparisons in the AEW (rigth) and intersect (left) datasets.* ***C.*** *The number of private alleles in the unfiltered (right above) and filtered (right under) AEW dataset.* ***D.*** *Folded site frequency spectrum (SFS) histograms.* ***E****. Two dimensional folded SFS spectra.*

differences in sample sizes. Watterson's estimates of theta (θ_W) ranged from 0.125% for the German population to 0.12%, 0.07%, and 0.05% for respectively the Aurignac, Ayrshire and East Anglia populations (Fig 3.6A). Nucleotide diversity (π) and genome wide heterozygosity (He) estimates ranged among populations between 0.04% and 0.16% (Fig 3.6), with π estimates being on average slightly below He estimates (Fig 3.6F). The mean He estimate in the German population was 0.12% (Fig. 3.6E-F), whereas the estimate obtained from a whole genome sequence analysis (of a sample derived from the same locality) equals 0.15% (this thesis, Chapter 4). This difference might suggest that our approach of estimating of He (and π) from RADseq datasets leads to an underestimate (for example due to missing data), or alternatively might represent a genome sampling bias.

Whereas θ_W estimates indicated that the German population harboured the highest genetic diversity, He and π estimates indicated instead that the Aurignac population was genetically the most diverse (Fig 3.6A, E-F). The difference between π estimates and θ_W estimates for the Aurignac population was reflected by a high Tajima's D score (Fig. 3.6A), and caused by an unsual high heterozygosity per segregating site (Fig 3.6D). This higher genetic diversity per segregating site outweighed the lower proportion of segregating sites, causing the nucleotide diversity of Aurignac to exceed the nucleotide diversity of the German population (Fig 3.6B,D,E,H). The Aurignac population did not contain population substructure (Fig. A3.8B), ruling out the Wahlund effect as potential explanation for the high genetic diversity within this population.

The East Anglia populations exhibited a signal typical for bottlenecked populations: reduced nucleotide diversity coupled with high proportions of common SNPs (Fig. 3.5D,E, 3.6G), indicating that many alleles, mostly of low frequency, were lost during and/or after the founder bottleneck. The Ayrshire and Aurignac population had different proportions of segregating sites (despite a roughly equal number of samples), but exhibited very similar site frequency spectra within those segregating sites. All populations except Wurttemberg (Germany) scored positive Tajima's D estimates, suggestive of a lack of rare alleles (Fig. 3.6A-B), potentially caused by population bottlenecks (and subsequent expansions).

Genetic diversity estimates of the Ayrshire population were intermediate to that of East Anglia and Wurttemberg (Germany) (Fig. 3.6D-F), indicating a similar

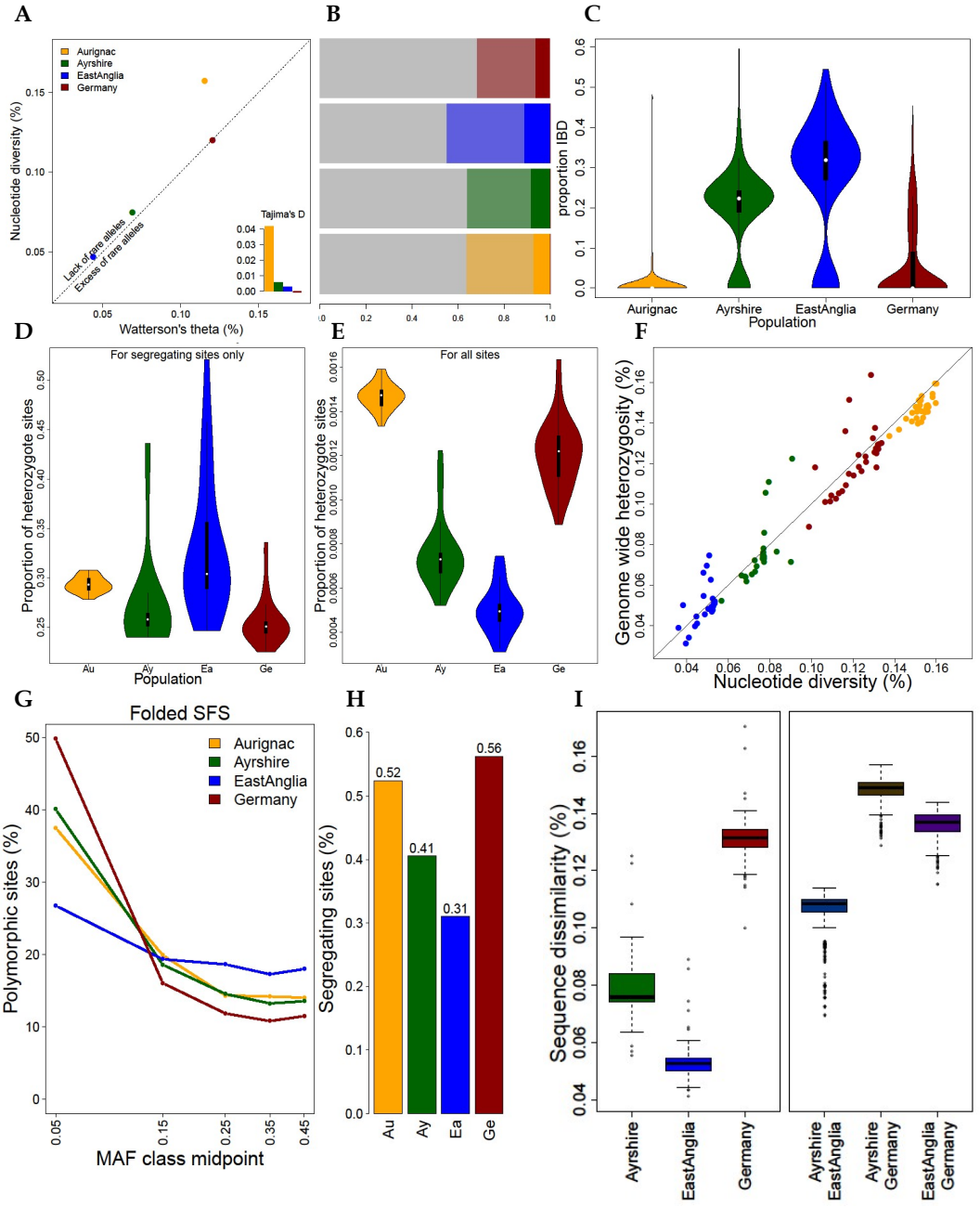


Fig. 3.6. Genetic diversity. *Colour coding: blue = EastAnglia (introduced UK), green = Ayrshire (native UK), red = Wurttemberg (Germany), orange = Aurignac (France).* ***A.*** *Watterson's theta, observed theta, and Tajima's D. All estimates are scaled per bp.* ***B.*** *Genotype proportions. Grey: major homozygous, light colour: heterozygous; dark colour: minor homozygous.* ***(C)*** *Proportion of genome identical by descent (pi_hat score, calculated with PLINK).* ***D.*** *Sample specific heterozygosity per segregating site.* ***E.*** *Sample specific genome wide heterozygosity.* ***F.*** *Sample specific gemome wide heterozygosity vs sample nucleotide diversity scores.* ***G.*** *Site frequency spectrum. Percentage of segregating sites per minor allele frequency class.* ***H.*** *Proportion of segregating sites.* ***I.*** *Sequence dissimilarity within and across populations.*

but less pronounced signal as observed for the East Anglia population: a loss of alleles, mostly of low frequency, likely due to genetic drift (Fig 3.6G). Although the SFS of Ayrshire was less distorted than the SFS of East Anglia (Fig 3.5Fig 3.9D), the Ayrshire population had a slightly higher Tajima's D score than the East Anglia population (Fig 3.6A), reflecting a stronger deviation of nucleotide diversity from Watterson's theta estimate, caused by differences in the distribution of the minor allele over minor homozygous and heterozygous genotypes (Fig. 3.6B).

***Demographic history*.** The stairway plot analyses for East Anglia identified a strong recent population bottleneck event, wrongly dated to around 1kya rather than 0.15kya. The stairway plot analysis furthermore pointed to a common size reduction in the other three populations between 10kya to 5 kya, with the Ayrshire and the Aurignac population being most heavily affected (Fig. 3.7). During that bottleneck, both the Ayrshire and Aurignac populations saw their Ne decrease from over 10k individuals to around 5K individuals. The historic effective population size of the Ayrshire population, which in this study represents the native UK population, is estimated to have been between 2,000 and 10,000 individuals, with a most likely value of 6,000 individuals (Fig 3.7).

Assuming a mutation rate of 1.1*10-8 per site per generation and a generation time of 5 years, the most likely onset of the Aurignanc population size decline seems to coincide with the end of the LGM (Fig 3.7). For Ayrshire, the most likely onset of the population decline seems to coincide with the end of the Younger Dryas (Fig 3.7). However, confidence intervals are wide, preventing exact timing of the population decline events (Fig 3.7).

Selection analyses. For all three pairwise comparisons (i.e. Ayrshire vs East Anglia, Ayrshire vs Germany, and East Anglia vs Germany), the distribution of locus specific Weir & Cockerham He-Fst estimates followed the same 'shark fin'-pattern as observed in Chapter 1 (Fig 3.8). The selection scans Fsthet, GWDS, and PCadapt did not detect outlier loci. PCadapt, in contrast, did mark a number of loci as outliers (Fig. 3.8). The Bonferroni corrected approach flagged up 16, 10 and 156 outliers for respectively the Ayrshire-East Anglia, the Ayrshire-Germany and the East Anglia-

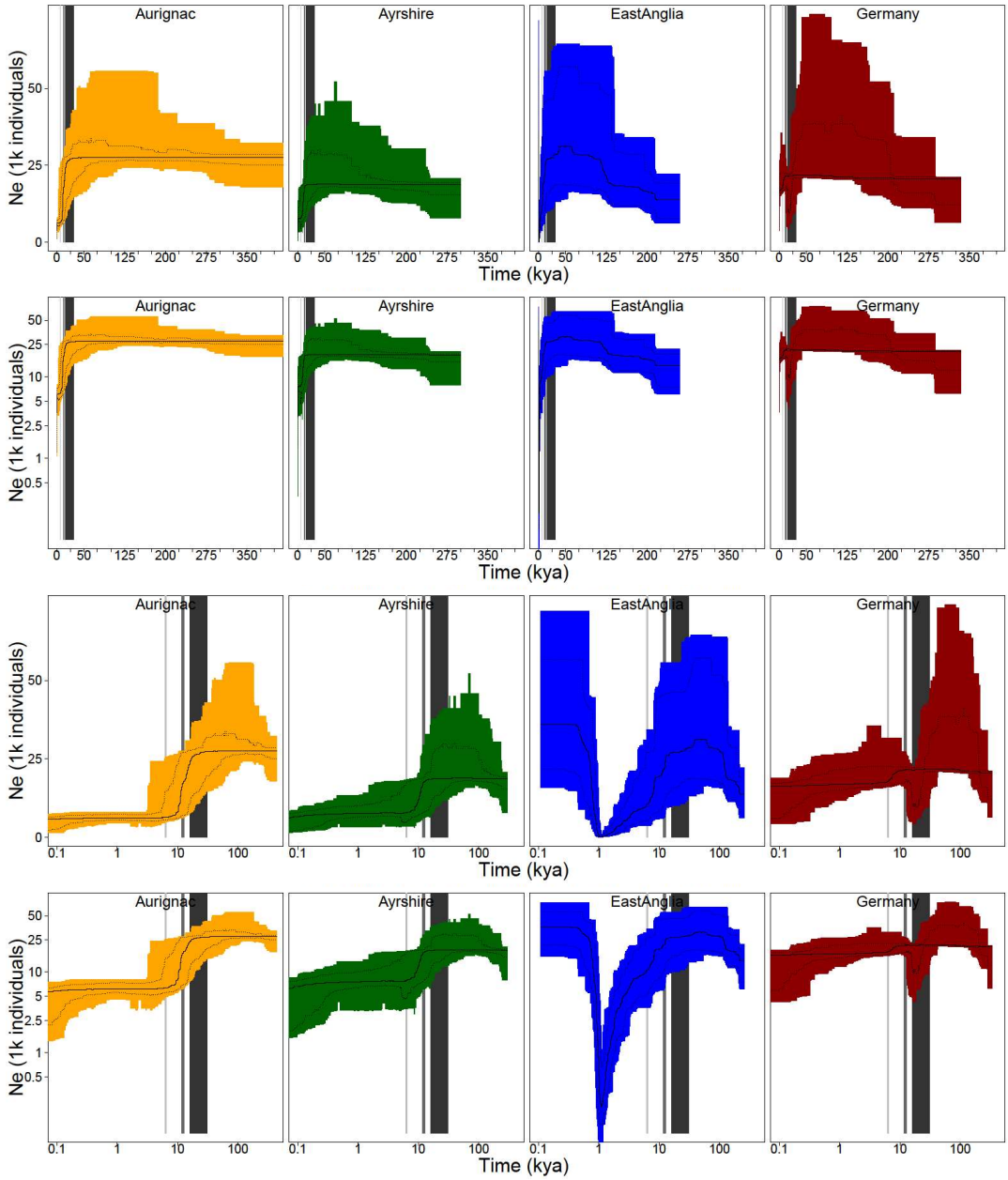


Fig. 3.7. Demographic histories. *Stairway plots showing demographic histories (mutation rate: 1.1e^{-8} per site per generation, generation time: 5 years), depicted using combinations of linear and log scales on the x- and y-axes. Grey shaded areas indicate from left to right: the flooding of Doggerland (~6 kya), the Younger Dryas (11.7-12.9 kya), and the last glacial maximum (16-31 kya). Solid lines indicate median values, whereas dashed line indicate 12.5% and 87.5% percentile values. Colour margins indicate 2.5% and 97.5% percentile values.*

Germany comparisons (Fig 3.8). My simulations indicated that given the demographic history of the native UK population, and assuming a selection coefficient of 0.01, most loci marked by PCadapt as outliers are false positives (Fig 3.9).

Similarly, running selection scans on a pooled comparison of modern UK populations (i.e. East Anglia and Ayrshire samples combined) vs the modern mainland (i.e. German) population, did not return outliers, except for PCadapt (Fig 3.8). In contrast, the pooled comparison of the native UK population vs the native mainland populations (i.e. East Anglia and Germany samples combined), flagged up two SNPs which were marked as outliers by both PCadapt and GWDS (Fig. 3.8). These two SNPs, which according to alignments to the *C. pygargus* genome occur alongside each other on contig 18718 on positions 1441634 and 1668556, both had a Weir & Cockerham Fst score of 0.85. Both SNPs had a minor allele frequency of 0.94 in the Ayrshire population and 0.05 and 0.1 in the East Anglian and German population (Fig 3.10).

Genes within 200kB distance of both outlier SNPs were genes coding for olfactory receptor 6C74-like protein, ras association domain containing protein 4, transmembrane protein 72, stromal cell-derived factor 1 protein, and two hypothetical, uncharacterized proteins (Fig 3.10). However, only one of the uncharacterized genes were however located within the outlier region (Fig 3.10). The other genes were separated from the outlier SNPs by multiple non-outlier SNPs (Fig 3.10).

Selection analyses control dataset with known selective sweep. The three selection scans (GWDS, OutFLANK and Pcadapt detected a locus under selection, signalled by multiple adjacent SNPs (Fig A3.10). The outlier region was detected for the pooled comparison as well as for both pairwise comparisons (Fig A3.10).

Discussion

In this study I harnessed the ddRADseq method to examine the degree and the causes of the genetic divergence of the native UK roe deer population, which got cut-off from European mainland populations due to Holocene sea level rise. In addition, I studied the historical demography and population structure of roe deer

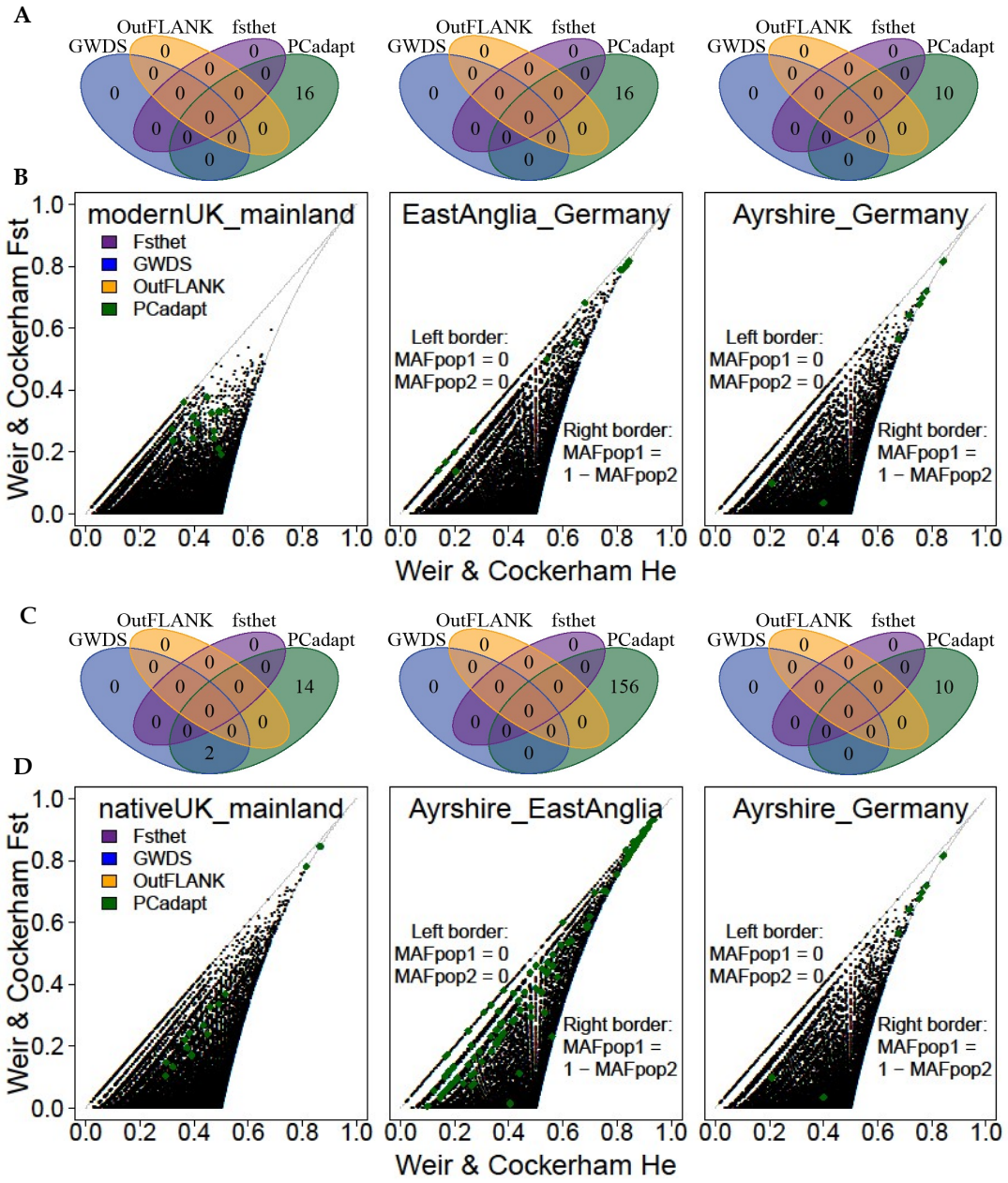


Fig. 3.8. Selection analyses. A-B. *Venn diagrams (A) and scatterplots of locus specific Weir and Cockerham heterozygosity and Fst values (B) depicting the number and overlap of outliers scored by selection scans Fsthet (purple), GWDS (blue), PCadapt (green) and OutFlank (yellow), for the comparison between modern UK populations (i.e. Ayrshire and East Anglia samples combined) and the modern mainland (i.e German) population..* ***C-D.*** *Venn diagrams (C) and scatterplots of locus specific Weir and Cockerham heterozygosity and Fst values (D) depicting the number and overlap of outliers scored by selection scans Fsthet (purple), GWDS (blue), PCadapt (green) and OutFlank (yellow), for the comparison between the ancient UK population (i.e. Ayrshire) and the ancient mainland populations (i.e samples from East Anglia and Germany combined).*

populations on either side of the North Sea, and assessed the impact of a population bottleneck on the genetic variation in an introduced population.

Ordination analyses identified the four study populations as distinct clusters and indicated that the Ayrshire population, which in this study represents the native UK population, is more closely related to the Wurttemberg (Germany) population than to the Aurignac (France) population (Fig. 3.2). This outcome appears to be in agreement with mitochondrial-DNA studies (Baker and Hoelzel, 2014; Fig. 2 in Randi et al., 2004) which indicated that after the LGM roe deer recolonized northwestern Europe and the British Isles from a central European lineage advancing through Germany, rather than from a southwestern European lineage advancing through France. Caution should however be exercised not to overinterpret these findings, because the ancestry of the Aurignac population is at present unclear, and also because the samples used in this study are derived from a limited number of populations which only partially represent European mainland populations.

The estimated genetic differentiation of the Ayrshire and Wurttemberg population (Fst: 0.123-0.14, Fig. 3.4A) is lower than the estimated differentiation of British and mainland European bank voles (*Myodes glareolus*; Fst: 0.229-0.358; Table S2 in Kotlík et al., 2018), but higher than the estimated differentiation of British and mainland European great tits (*Parus major*; Fst: 0.003-0.006, Fig. 1 in Bosse et al., 2017). These among species differences in observed Fst-values is likely partly accounted for by species traits, most specifically the combination of generation time and effective population size (Ne). Great tits have exceptionally high effective population sizes (i.e. Ne > 500,000 individuals, Fig 1B in Laine et al., 2016) which minimalizes genetic drift. Given the differences in body size, the Ne of bank voles will also likely be higher than the Ne of roe deer, but possibly not as high as those of great tits. The shorter generation time of bank voles, in addition to an earlier establishment in the UK (Searle et al., 2009), might explain why British bank voles are genetically more diverged from their mainland counterparts than British roe deer are from their mainland counterparts.

The effect of Ne on genetic divergence, through the workings of drift, is illustrated by the bottlenecked East Anglian population. Although the East Anglia population split from the Wurttemberg (i.e. German) population less than 150 ya,

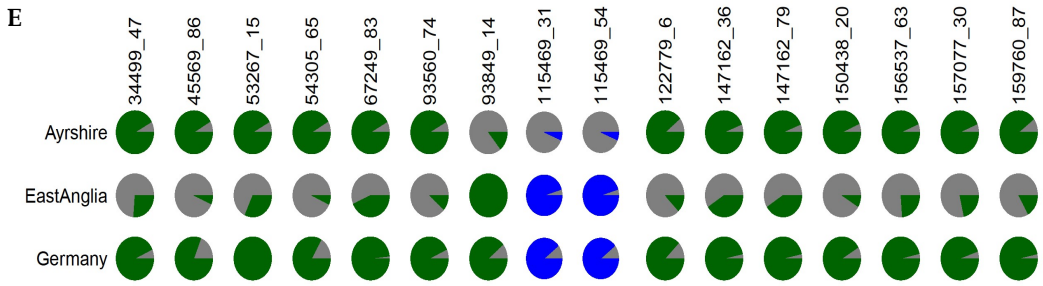
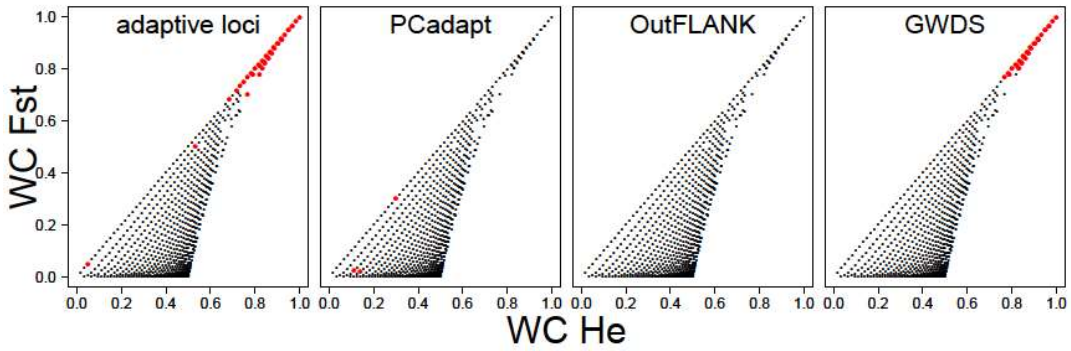


Fig. 3.8. Selection analyses cont. E. *Piecharts of allele frequencies in each of the populations. Green piechart indicates SNPs marked by PCadapt as outlier loci for the native UK vs mainland comparison. Blue: SNPs which are marked as outliers by both PCadapt and GWDS. The allele frequencies of the two outlier SNPs which are marked by both PCadapt and GWDS are 0.94 in the Ayrshire population and 0.05 and 0.10 in the East Anglian and German population.*



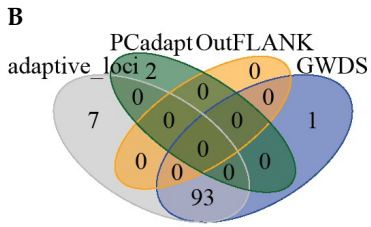


Fig. 3.9. Detectability of loci under diversifying selection according to simulations. *Simulation output showing the detectability of 100 SNPs under diversifying selection (out of 10,000 SNPs in total) for pairwise population comparison. Demographic scenario: mainland roe deer population Ne = 10,000, native UK roe deer population Ne = 5000, TMRCA = 1500 generation, with a sample size of 30 individuals per population. It is assumed that the ancestral population was panmictic (i.e. no isolation by distance).* ***A.*** *Simulated He-Fst distributions for pairwise population comparison. Black: 10,000 neutral loci. Red: 100 loci under weak diversifying selection (s=0.01) (left panel) or loci marked as outliers by the selection scans PCadapt, OutFLANK and GWDS (other panels)* ***B.*** *Venn diagram showing the number of simulated adaptive loci (out of 100 in total) correctly marked as outliers by PCadapt, OutFLANK and GWDS in the pairwise population comparison.*

this population is more differentiated (i.e. Fst = 0.19-0.224, Fig. 3.4A) from the German population than the native UK population, which has been separated for over 6000y. Sequence dissimilarity estimates convey a different message (i.e. lower dissimilarity scores for EastAnglia-Germany than for Ayrshire-Germany, Fig. 3.6I), but this is likely due to the increased loss of low frequency alleles within the heavily bottlenecked East Anglia population, which increases sequence similarity between East Anglia samples and the majority of German samples.

Landbridge island populations, in particular those occuring on smaller islands, typically contain less genetic variation than closely related mainland populations (Bell et al., 2012; Hurston et al., 2009; Lourenço et al., 2018; Robinson et al., 2016; Velo-Antón et al., 2012; Wang et al., 2014). In agreement with findings based on mt-DNA comparisons (Baker and Hoelzel, 2014), I found that the native UK population (i.e. Ayrshire population) harbours less genetic diversity than the central European roe deer lineage (i.e. German population). In contrast, genomic studies on great tits (Bosse et al., 2017) and bank voles (Kotlík et al., 2018) do not indicate marked lower genetic variation of UK populations compared to European mainland populations, and neither do mitochondrial and microsatellite-DNA studies on red fox (*Vulpes vulpes*, Atterby et al., 2015; Edwards et al., 2012) and badgers (*Meles meles*, Fig S1 in Frantz et al., 2014). The comparatively low genetic diversity of the native UK roe deer population is therefore in need of explanation. This explanation might be anthropogenic influence, but overhunting during medieval affected the English roe deer population in particular (Baker and Rus Hoelzel, 2012).

A well established signature of drift is a negative relationship between genetic diversity and divergence, with the least genetically diverse populations being most diverged from the ancestral population (Funk et al., 2016). Consistent with this expectation, I found that the East Anglia population has lower nucleotide diversity than the Ayrshire population. Some microsatellite DNA studies suggest that bottlenecks (i.e. founder sampling) affect allelic diversity more than they affect heterozygosity (Lampert et al., 2007). As observed and discussed in Chapter 2 of this thesis as well, I found that, probably likely due to the loss of low frequency alleles, individuals in the bottlenecked East Anglia population contained higher heterozygosity for segregating sites than individuals from non-bottlenecked

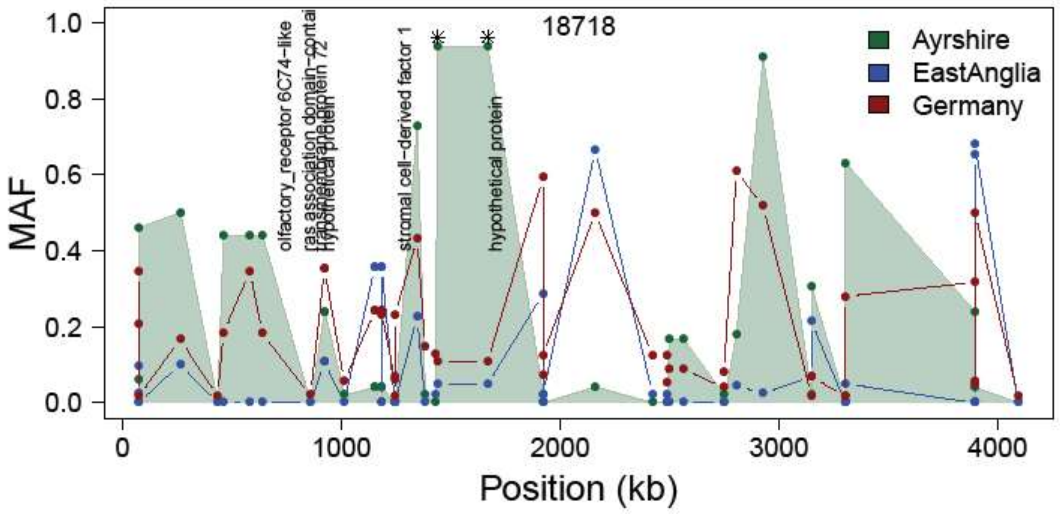


Fig.3.10. Genes close to outlier SNPs. *Genes within 200kB distance of the two SNPs marked as outliers in the nativeUK vs mainland comparison by both GWDS and PCadapt, according to alignment to the C. pygargus genome (see Chapter 4 of this thesis). Shading and lines show population specific minor allele frequencies of each SNP. Outlier SNPs are indicated with an asterix. Detected genes are olfactory receptor 6C74-like, ras association domain containing protein 4, transmembrane protein 72, stromal cell-derived factor 1, and two hypothetical, uncharacterized proteins.*

populations do (Fig 3.6D). The picture reverses when heterozygosity is averaged over all sites (i.e. both segregating and non-segregating sites, Fig 3.6E), as this estimate also takes into account the loss of these low frequency alleles.

The loss of low frequency alleles within the East Anglia population is reflected by a flat site frequency spectrum (SFS, Fig 3.5B, 3.6G). Stairway plot analysis correctly infers a recent population bottleneck (Fig 3.7) from this SFS, which serves as a proof of method. The Stairwayplot analyses on the other study populations indicate that the Ayrshire, Aurignac and Germany population experienced a shared population size reduction at the start of the Holocene (Fig 3.6). The timing of this event is imprecise (Fig 3.6) and depends on settings (i.e. mutation rate and generation time), but potentially coincides with a period of rapid warming following the Younger Dryas.

The fossil records suggest that roe deer were absent north of the Alps during the Younger Dryas (Sommer et al., 2008). One possible scenario is that the common reduction in population size reflects a founder effect caused by range expansions (Eckert et al., 2008), more specifically the recolonization of northwestern Europe in a period of rapid warming following the Younger Dryas.

This scenario would explain why the German population is less affected by the bottleneck than the Ayrshire population, as the German sampling area was located closer to roe deer refugia during the Younger Dryas (Sommer et al., 2009). According to this interpretation the lower levels of genetic diversity within the Ayrshire population are due to natural causes, and not to antrophogenic events. An alternative explanation of the observed population declines is a shared response to an environmental driver.

The distributions of locus specific He-Fst estimates for pairwise population comparisons (i.e. Ayrshire vs Germany, East Anglia vs Germany, and Ayrshire vs East Anglia) resembles a 'shark fin' which was also reported in Chapter 2 of this thesis and in a study by Zucchi et al.,(2019). Flanagan et al (2017) have shown that this type of distribution can not confidently be screened by Fdist selection scan methods such as Lositan (Antao et al., 2008) and Fdist2 (Excoffier and Lischer, 2010). Fdist methods assume an island model with potentially ongoing gene flow between populations. Absence of gene flow – as in the case for roe deer populations occurring on either side of the North Sea – may lead to discrepancy of the expected and observed distribution of locus-specific He-Fst values (Fig 1B and Fig. 3 in Flanagan and Jones, 2017) and consequently to high false positive rates.

Both observed and simulated He-Fst distributions illustrate that given a split time of ~6ky (flooding of Doggerland) and a generation time of 4 years, genetic drift alone is not sufficient to drive segregating alleles, let alone newly derived alleles, to fixation. Whereas genetic drift causes populations to diverge slowly (Watterson, 1975), selection can cause fixation of adaptive alleles within a few hunderd generations, depending on the magnitude of the selection coefficient (Kimura and Ohta, 1969). Although I didn't find fixed differences between the native UK population (i.e. Ayrshire) population and the native mainland (i.e German) population (Fig 3.5E), the selection scans did mark two adjacent SNPs with near fixed differences as outliers. This SNPs had a minor allele frequency of 0.94 in the Ayrshire population and of 0.05 and 0.1 in the East Anglian and German population (Fig 3.10) and were highlighted by both GWDS and PCadapt in the native UK vs native mainland comparison (Fig 3.8).

In comparison to the outlier locus detected (Fig A3.10) for the control human chromosome-2 SNP dataset (known to harbour a locus responsible for lactose

tolerance in north-western European populations), the evidence for a positive selection event in the native UK population appears weak, for two reasons. First, the outlier locus in the human dataset is signalled by >30 out of 80.000 SNPs, whereas the outlier locus in the roe deer dataset is signalled by two SNPs only. This difference can (partly or wholly) be attributed to the difference in density of the SNP catalogue (on average 1 SNP per ~2Kb for the human dataset vs 1 SNP per ~40 Kb for the roe deer dataset). Second, the outlier SNPs in the human dataset differ more strongly from the neutral distribution (i.e. higher selection scan test scores and higher difference in Fst-values) than the outlier SNPs in the roe deer dataset. Due to the relatively wide neutral Fst-distribution of the roe deer dataset, outlier SNPs have less potential to stand out from the neutral distribution. The inflated neutral distribution, which leads to the relatively low test scores of the roe deer SNP outliers, therefore does not allow to rule out that the SNP outliers are false positives caused by a stochastic abberation of drift affecting one particular locus more strongly than other loci. This effect highlights the limited applicability of Fst-outlier tests, which in the absence of gene flow loose power if the TMCRA approaches 4·Ne generations.

To my knowledge, this study is the second study to present potential evidence for outlier regions possibly under diversifying selection between UK and European mainland populations, inferred from SNP datasets. Previously, a high density SNP catalogue revealed several putatitive outlier genomic regions under anthropogenic diversifying selection between British and Dutch populations of great tits (Fig S3A,B in (Bosse et al., 2017). Locus specific Fst values of SNPs in these outlier regions were at maximum 0.15, which is much higher than genome wide averages (Fst = 0.006, Bosse et al., 2017) but also seems to indicate that the adaptive alleles are still segregating (i.e. no fixed differences).

Ample evidence for post-LGM diversifying selection is found in post-glacial lakes and seas, which – as famously illustrated by the threespine stickleback – are often home to various ecotypes despite the lack of obvious geographical boundaries which could limit gene flow (Hohenlohe et al., 2010; Schluter et al., 2010). Genome wide selection analysis resulted in 48 (1.22%) out of 3925 SNPS being highlighted as being possibly under diversifying selection between two morphologically and ecologically differentiated ecotypes of trout occurring in a post-glacial lakes in

Canada (Bernatchez et al., 2016). Arlequin, Bayescan and OutFLANK detected 8 out of 2,051 SNPs (0.39%) as divergent between pelagic and demersal spawning European flounders in the postglacial Baltic Sea (Momigliano et al., 2017). None of the studies reported outliers characterized by fixed differences.

One of the most studied phenotypic differences between insular and mainland populations are differences in body size (Losos and Ricklefs, 2009). These body size differences have been argued to be driven by abiotic factors, particularly community structure (Keogh et al., 2005; Lomolino et al., 2013). The extent of dwarfism in ungulates depends on the existence of competitors and to a lesser extent on the presence of predators. In carnivores, body size has been found to be associated with prey abundance and prey size (Raia and Meiri, 2006). The theory of island biogeography predicts that due to the dependency of migration and extinction probabilities on island size, smaller islands contain less species (Itescu et al., 2019; MacArthur and Wilson, 2001) and therefore that differences in community structure between islands and mainland, and hence selective pressures on body size, depend on island size. Measuring over 200,000 km^2, Great Britain is among the biggest islands worldwide, and consequently the faunal composition of Great Britain is very similar to the faunal composition of north western Europe (Montgomery et al., 2014; Stuart, 1995). This faunal similarity might equate to the absence of biotic diversifying selection.

The fact that the two outlier SNPs are adjacent – mirroring results presented in Chapter 2 of this thesis – and furthermore have identical genotype scores, makes it highly unlikely that that these SNPs stand out due to genotyping errors. As I did not detect any known genes or other genomic features within the outlier region, any inferences about the exact nature of the selective event are purely speculative. Given the faunal similarity between Britain and the European mainland, it seems reasonable to assume that the selective driver is abiotic. Islands and adjacent mainlands are environmentally and climatically highly heterogeneous (Weigelt et al., 2013). The British Isles have an unique climate, and it has for example been hypothesized that the distinct morphology of the Irish bee aids survival in the damp cool climate of Ireland (Hassett et al., 2018). At the same time, due to the size of Great Britain, many climatic factors differ within Great Britain as much as they differ between Great Britain and the mainland. The outlier region might therefore

represent local adaptation instead of a British or mainland adaptation, but additional sampling across Scotland would be needed to exclude either scenario.

Conclusions

In this study I provide evidence that the Ne of the British roe deer population has numbered several thousand throughout the Holocene, resulting in moderate levels of genetic drift which have led to moderate loss of standing genetic variation. Based on comparisons of the study populations (i.e. Ayrshire, Aurignac and Wurttemberg populations), genetic diversity within the native British roe deer population falls below the genetic diversity of the mainland roe deer population. Selection scans identified 2 adjacent outlier SNPs out of over 50K SNPs in total. The genomic region in which these SNPs occur potentially experienced diversifying selection in either the native UK or the mainland roe deer population, possibly associated with climatic differences.

Chapter 4

Demographic and evolutionary history of roe deer sister species (*Capreolus* spp) inferred from whole genome sequencing data

Abstract

Species that evolved during the Pleistocene in temperate regions experienced periods of extreme climatic transitions, but it is still unclear how these climatic events impacted their evolutionary histories. The parapatric distribution of the two extant roe deer species, the European roe deer (*C. capreolus*) and the Siberian roe deer (*C. pygargus*), suggests secondary contact following allopatric speciation, possibly facilitated by climatic transitions. Here I make use of a new high-coverage reference genome for *C. pygargus* in combination with publicly available deer genomes, including the low quality reference genome of *C. capreolus*, to infer the demographic and evolutionary history of extant roe deer. My analyses suggest a more recent split time (≤1.6Mya) of the *Capreolus* sister species than previously suggested by mtDNA studies (~2-4Mya), pronounced differences in terms of their genetic diversity and effective population sizes, and contrasting demographic trajectories. In the species with lower genetic diversity and lower historical Ne estimates, *C. capreolus*, I find higher proportions of lineage specific amino acid substitutions. I hypothesize that these elevated dN/dS rates in *C. capreolus* reflect episodic positive selection events, enhanced by low effectiveness of puryfing selection typical for small populations. In conclusion, I suggest that both selective and neutral processes have influenced the divergence of the two sister taxa.

Related peer-reviewed publication:

De Jong, M.J., Li., Z., Qin, Y., Quemere, E., Baker, K., Wang, W. 2020. *Demography and adaptation promoting evolutionary transitions in a mammalian genus that diversified during the Pleistocene*, Molecular Ecology

Author contributions:

ARH conceived the study and MdJ & ARH wrote the paper. MdJ undertook data and lab analyses. EQ generated the RADseq data for the Aurignac population. ZL, YQ and WW generated the C. pygargus genome assembly and annotation.

Introduction

The climatic oscillations in the Pleistocene (2.59–0.01Mya) serve as natural experiments which provide insights into the evolution of populations in the face of rapidly changing environmental conditions (Hofreiter and Stewart, 2009). One major finding is niche conservatism. Populations predominantly respond to changing environmental conditions by habitat tracking (and/or phenological shifts) rather than genetic tracking, resulting in tidal-like fluctuations of range limits (Hewitt, 2000, 2004; Nadachowska-Brzyska et al., 2015; Stewart et al., 2010)

When environments change, niche conservatism can cause fragmentation of populations and hence facilitate allopatric speciation (Avise et al., 1998; Wiens, 2004). Pleistocene climatic oscillations have therefore been hypothesized to drive speciation, both in temperate and non-temperate regions (Haffer, 1969; Klicka and Zink, 1997). But despite the increased potential for population fragmentation, Pleistocene speciation rates do no stand out from other geological era, not do they exhibit pulses correlated with climatic transitions, suggesting that speciation is neither facilitated nor inhibited by the glaciation cycles (Barnosky, 2005; Bibi and Kiessling, 2015; Klicka and Zink, 1997, 1999; Lister, 2004).

These ordinary and continuous speciation rates in an era of increased climatic instability can be seen as evidence favouring the hypothesis that evolutionary change is driven by biotic interactions rather than by abiotic factors such as climatic change (Benton, 2009). An alternative explanation is however that populations generally need to be isolated for longer than the typical duration of glacial-interglacial cycles in order to complete the speciation process (Barnosky, 2005).

In this study I performed comparative genomic analyses of two mammalian sister species which evolved during the Pleistocene: the European/western roe deer (*Capreolus capreolus*) and the Siberian/eastern roe deer (*Capreolus pygargus*). These two sister species are phenotypically very similar, *C. pygargus* being bigger and bearing greater and more branched antlers (table 1 in Plakhina et al, 2014). A large part of the morphological, ethological and ecological variability of roe deer can be contributed to within species differences rather than between species differences (Danilkin, 1995).

On the genetic level European roe deer have a fixed chromosome number (2n=70) whereas Siberian roe deer have various chromosome numbers (2n=70+(2B/4B/14B), Xiao et al., 2007). Mitochondrial DNA control region studies have indicated that the range of pairwise sequence dissimilarity between individuals of either species ranges around 4.9-5.8% (Randi et al., 1998; Xiao et al., 2007). These estimates lies firmly within the range reported for other deer species pairs (i.e. 4.7% to 6.9%; (Douzery and Randi, 1997, cited in Xiao et al., 2007) and contrasts with the pairwise sequence dissimilarity between individuals within species, which ranges below 3.0% (Xiao et al., 2007). Further evidence for the species status of both roe deer types comes from the observation that most captivity born hybrid males are sterile (Sokolov and Gromov, 1990).

Based on their mtDNA control region differentiation the two species are thought to have diverged between 2 to 3.7 mya (i.e. 2.2-3.7 mya according to Douzery and Randi, 1997) and 2-3 mya according to Randi et al. (1998); both estimates cited in Xiao et al. (2007). At present the two species maintain a parapatric distribution in the temperate zone of the Eurasian continent and share a border which runs in longitudinal direction through southwestern Russia (Fig. 1). The hybrization zone surrounding this border is thought to extend from the right side of the Volga river up to Eastern Poland (Plakhina et al., 2014). The border between the two species lacks obvious geographical boundaries which could limit gene flow, and does not overlap with obvious environmental boundaries. The location and orientation of the species border, as well as the sizes of both species distribution ranges, suggest that environmental variables vary more within than between ranges.

In theory, parapatric distributions of sister species can originate through either speciation with ongoing gene flow (Martin et al., 2013; Morales et al., 2017; Wang et al., 2019; Winker et al., 2019) or through secondary contact following allopatric speciation (Pastene et al., 2007; Poelstra et al., 2014). The first scenario entails divergence through diversifying selection in heterogeneous environments despite the homogenizing effect of gene flow. Binary phenotypic divergence along environmental gradients can potentially result from a threshold response, an abrupt change in favoured phenotype along the gradient (Riesch et al., 2018).

The second scenario entails a three step process of range fragmentation due to a vicariance event, divergence in isolation, and range reunion. This three step process potentially facilitates a non-adaptive diversification event, in which diversification of a lineage is not accompanied by relevant niche differentiation (Comes et al., 2008; Gittenberger, 1991; Lambert et al., 2019). Non-adaptive diversification can be driven by either mutation-order speciation (i.e. selective driven fixation of mutations; (Czekanski-Moir and Rundell, 2019; Schluter, 2009) or by neutral speciation (i.e. fixation of new mutations and/or standing variation by drift (Orr and Orr, 1996).

Given the apparant absence of relevant niche differentiation between *C. capreolus* and *C. pygargus*, it seems plausible that the current parapatric distribution of the two extant roe deer species is a vestige of a diversification event driven by climate change induced range fragmentation and subsequent reunion. Unknown is however whether this diversification event was driven by mutation-order speciation or by neutral speciation.

In this study I compared a new, high quality genome assembly of the Siberian roe deer (*C. pygargus*) to available genomes of other cervid species, including the low quality genome assembly of it's sister species, the European roe deer (*C. capreolus*). My objectives were twofold. My first objective was to gain more insight in the demographic history of *C. capreolus* and *C. pygargus* through estimating their historical Ne and their TMRCA. My second objective was to assess to what extent the divergence of *C. capreolus* and *C. pygargus* has been driven by diversifying selection, which could reflect mutation-order speciation. To that end I searched for genes which experienced episodic positive selection in either sister species using PAML's codeml as well as custom-built tool to detect accelerated dN/dS rates within foreground lineages.

Although PAML's codeml is a very popular method to search for adaptive substitutions within genes, only a minor subset of studies has screened full exomes for species specific adaptive substitutions (rather than clade specific adaptive substitutions), the reason being that many whole genome sequences have only recently become available. For studies which do not contain multiple species per genus or per subfamily, it is not clear whether outlier genes reflect episodic selection in the lineage leading to the species, or earlier episodic selection on earlier lineages.

In this study I compared exomes of >10 Cervidea species, including *C. pygargus*, *C. capreolus* and the sister lineage *H. inermus*, allowing us to differentiate between genus specific selective events and species specific selective events.

Methods

Acquisition of raw reads and genome assembly of *C. capreolus*. The raw reads and genome assembly of *C. capreolus* were generated for a previous study (Kropatsch et al., 2013) and kindly provided to us by the authors. The authors collected a blood sample from a male roe deer from Hohenstein-Born (Germany; 50'09'' N, 8'05'' E) and prepared this sample for paired end sequencing on an Illumina 1.9 platform. The full details of the sequencing protocol are described in Kropatsch et al (Kropatsch et al., 2013). The *C. capreolus* assembly has a total length of 2,785,377,831 bp, distributed over 314,210 scaffolds (of 1kb or longer), with a median (N50) scaffold length of 10,458 bp.

Of the 422,979,622 + 422,818,638 *C. capreolus* forward and reverse reads, I dropped respectively 142,876 and 124,768 reads which were contaminated with adapter sequences, retaining 422,836,746 forward and 422,693,870 reverse reads. I used the software Trimmomatic (Bolger et al., 2014) to discard all *Capreolus capreolus* reads with an average PHRED33-quality score below 20, retaining 412,716,619 read pairs. All reads were trimmed to a length of 101 bp.

Acquisition of raw reads and genome assembly of *C. pygargus*. The raw reads and genome assembly of *C. pygargus* were generated by the Center for Ecological and Environmental Sciences of the Northwestern Polytechnical University in cooperation with the Department of Special Animal nutrition and Feed Science of the Institute of Special Animal and Plant Sciences of the Chinese Academy of Agricultural Sciences, and kindly provided to us before publication. DNA was extracted from liver tissue of a 7 month old male roe deer from the Er He wild animal farm at Shulan, Jilin city, Jilin province, in northeastern China (126'58'' N, 43'855'' E), and subsequently prepared for 10X Genomic Chromium system sequencing. The *C. pygargus* assembly has a total length of 2,607,875,777 bp, distributed over 92,100 scaffolds, with a median (N50) scaffold length of 6,607,211 bp.

Genome wide heterozygosity. I used Bowtie version 2.2.5 to map the retained sequence reads of both species to their reference genomes. I used Samtools version 1.3.3 to filter out reads with a mapping quality below 20 and applied the command 'grep –v 'XS:i'' to filter out reads which mapped to multiple locations, retaining 676,985,798 *C. pygargus* and 522,854,805 *C. capreolus* reads. I subsequently called SNPs using samtools, bcftools (Narasimhan et al., 2016) and vcftools (Danecek et al., 2011), and used tcsh command line tools to count on a per contig basis the number of heterozygous sites, the total number of sites with genotype information and the spacing between adjacent heterozygous sites. For comparison I used the same approach to generate He estimates for two other deer species, namely white tailed deer (*O. virginianus*) and red deer (*C. elaphus*).

Average read depths after filtering, calculated using the samtools depth tool, equalled 22.1 for *C. capreolus* and 39.7 for *C. pygargus.* To investigate the dependency of genetic diversity estimates on average read depth, I randomly downsampled the *C. pygargus* bam file, using the samtools view tool with the -s flag set to 0.53. The value of s was derived using the following formula: (522,854,805 *C. capreolus* reads x 101 bp per *C. capreolus* read)/(676,985,798 *C. pygargus* reads x 150 bp per *C. pygargus* read).

Runs of homozygosity. I used two methods to screen the *C. capreolus* and *C. pygargus* genomes for runs of homozygosity (ROHs). The first method was based on the distance between adjacent SNPs. Because of the low contig sizes for the *C. capreolus* genome, I used the *C. pygargus* genome as the reference genome for both species, assuming highly conserved synteny. The assumption of synteny among cervids was verified with dot plots (Fig. A4.1), which I generated by mapping all *C. pygargus* contigs of 10 Mb or longer to *C. elaphus* chromosomes using the software Lastz version 1.02.00 (Harris, 2007) using the 'gfextend', 'chain' and 'gapped' –options.

He-spacing statistics were calculated as the distance between heterozygous sites. Inter-He-regions which were truncated at the start or the end of a contig (and hence were flanked by one rather than two heterozygous sites), were included in the analysis. After calculation of the distance between heterozygous sites, a sliding window approach was used to screen the *C. capreolus* and *C. pygargus* genomes for

regions with above average spacing between heterozygous sites (i.e. genomic regions with depleted genetic variation), using the rollapply function of the R-package 'zoo' (Zeileis, 2005). I set both stepsize and windowsize to 100 datapoints (i.e. 100 adjacents inter-He-regions). I excluded from the analysis all inter-He-regions with 10 percent or more missing data points.

A window was marked as an outlier window if it met two criteria: a) the window should contain at least 1 inter-He-region in the top 0.05% (5%/windowsize) of all regions; and b) the window should contain at least n regions in the top 5% of all regions located on the respective contig, with n averaging 15, the exact number being dependent on the number of windows (nwin) on the contig, as described by the following R function qpois: n = qpois((1-0.05/nwin), (windowsize/20)). Contigs with a size below 5 Mb were not considered, retaining 164 contigs with a median and mean length of respectively 8.0 Mb and 9.3 Mb (sd = 4.4 Mb) and a combined length of 1519.4 Mb, spanning roughly half of the genome.

The second approach used to detect ROHs involved calculation of heterozygosity on a sliding window basis using various window sizes, ranging from 10Kb to 3Mb, and using non-overlapping windows (i.e. step size equalled window size). These estimates were generated using a combination of windows command line tools, as well as the software tabix (Li, 2011) to subselect vcf files. R command line tools were subsequently used to calculate the proportion of windows with a heterozygosity below a specified threshold (namely 0.1%, 0.05% and 0.01%). F_{ROH} was defined as the total length of windows below the He-threshold, divided by the total length of all windows, excluding missing data points. Contigs shorted than 10Mb, and windows with more than 20 percent data, were excluded from the analyses.

PSMC analyses. I generated a diploid fasta file of both genomes by mapping the raw reads of both species to their respective genomes, calling snps using samtools mpileup and bcftools, and by converting to fastq files using the vcfutils.pl executable of bcftools. To correct for differences in read depth between datasets, I downsampled the *C. pygargus* data to match the read depth of *C. capreolus*.

Sequentially Markov coalescent modelling (McVean and Cardin, 2005) was executed using the software PSMC (Li and Durbin, 2011), with the default settings

of 64 time intervals defined by 28 parameters (-p "4+25*2+4+6"), and a maximum number of 25 iterations (-N25). I also used the default settings of -t15 and -r5. Maximum read depth was set to twice the average read depth (i.e. D = 42), and minimum read depth was set to one third the average read depth (i.e. d = 7). A minimum read depth of 7 is slightly below the minimum read depth of 10 recommended by Nadachowska-Brzyska et al (2016). The mean read depth of 21x is, in contrast, above the recommended coverage of 18x (Nadachowska-Brzyska et al., 2016). The *C. capreolus* assembly has limited contiguity, with a reported scaffold N50 of 10.458 bp (Kropatsch et al., 2013). Simulation analyses have shown that PSMC analyses are relatively robust for scaffold sizes down to 10kb, depending on the demographic history (figure 1 in Chapter 3 of Gower, 2019).

I set the generation time parameter to 4-6 years (Nilsen et al., 2009). I assumed a mammalian mutation rate per site per year of $0.22 \cdot 10^{-8}$ (Fig. S29 in Chen et al., 2019; Kumar and Subramanian, 2002) and, assuming a linear relation, a mutation rate per site per generation (5 years) of $1.1 \cdot 10^{-8}$. For bootstrapping I used 100 replicates.

Genome wide genetic divergence. I calculated sequence (dis)similarity between both *Capreolus* sister species by crossmapping raw reads to whole genome sequences (i.e. *C. capreolus* reads to *C. pygargus* genome, and *C. pygargus* reads to *C. capreolus* genome) using Bowtie2, and subsequently calling SNPs and indels using samtools, bcftools and vcftools. I filtered out reads with a mapping quality below 20 and applied the command 'grep –v 'XS:i'' to filter out reads which mapped to more than one location, as well as sites with a read depth below 8. I counted the total number of sites and SNVs on a per contig basis using tcsh command line tools. Sequence dissimilarity was estimated as the proportion of fixed differences plus half the proportion of segregating sites.

Split time estimation using a random walk Markov chain model. I calculated the TMRCA by estimating the time (in years or generations) needed to obtain the observed genome-wide pairwise sequence dissimilarity. I estimated the duration of this time interval by applying a custom-built random walk Markov chain model in

which I simulated the proportion of single nucleotide differences between two sister taxa after a vicariance event.

For simplification I assumed that a mutation will fixate instantly, within a single generation, meaning that substitution equals mutation, and that a site is always fixed (i.e. no segregating sites). In reality it will take a newly arisen allele on average 4Ne generations to fixate within a diploid population (Kimura and Ohta, 1969), suggesting that the TMCRA estimates generated by my Markov chain method underestimates the true TMCRA. For simplicity I also assumed equal substitution rates, and equal mutation rates between pyrimidines and purines (i.e. the model assumes transversion rates to equal transition rates). The model assumes the absence of admixture (i.e. no gene flow), but does take into account the affect of incomplete lineage sorting, by assuming random fixation or loss of the standing variation within either sister taxa.

Let the symbol 'u' denote the probability of a point mutation per site per generation. For each moment in time (and therefore independent of the number of generations since the vicariance event) I can make the following argument: If for a given locus both taxa have the same DNA base (S for Similar), then the probability that in the next generation they will differ for that particular locus, is the sum of two probabilities:

- the probability that one taxon experiences a mutation and the other does not: $2u(1-u)$
- the probability that both taxa have a mutation, but to different bases: $(1/3)u^2$.

Combined probability = $2u - (5/3)u^2$.

If for a given locus both taxa have a different DNA base (D for Dissimilar), then the chance that in the next generation the taxa will be similar for that particular locus, is again the sum of two probabilities:

- the probability that one taxon mutates towards the other taxon (so: one taxon mutates, the other doesn't): $2(1/3)u(1-u)$
- the probability that both taxa happen to mutate to the same DNA letter: $(1/2)u^2$.

Combined probability = $(2/3)u - (1/3)u^2$.

This Markovian model, which consists of two states (S and D) and two transition probabilities ($Pr_{S=>D} = 2u - (5/3)u^2$ and $Pr_{D=>S} = (2/3)u - (1/3)u^2$), can be described by the following recursive formula:

$S(n+1) = S(n) + ((2/3)u - (1/3)u^2)(1 - S(n)) - (2u - (5/3)u^2)S(n)$

Now, let C denote $((2/3)u - (1/3)u^2)$ and let k denote $(2u - (5/3)u^2)$:

$S(n+1) = S(n) + C(1-S(n)) - kS(n)$

This can be rewritten to:

$S(n+1) = (1 - k - C)(S(n) + C)$

Now let r denote $(1 - (k - C))$:

$S(n+1) = rS(n) + C$

This can be rewritten to the following decay function:

$S(n) = S(0)r^n + C(r^n-1)/(r-1)$

in which:

r = $1 - (8/3)u + 2u^2$

C = $(2/3)u - (1/3)u^2$

S(0) = the initial value of the similarity of the sister taxa after random fixation or loss of standing variation (approximated by theta of ancestral population)

n = number of years/generations

u = mutation rate per site per years/generation

I solved this formula for sequence similarity estimates derived from cross mapping raw reads to the genomes of the sister species. I assumed a predefined mutation rate (u) of $1.1 \cdot 10^{-8}$ per site per generation and $0.22*10^{-8}$ per site per year (Fig S29 in Chen et al., 2019; Kumar et al., 2015). I also ran forward time simulations to derive 95% confidence intervals. I tested the validity of the model by comparing predicted TMRCA estimates with published estimates on great ape divergence times and sequence dissimilarity.

Gene alignment and exome species trees. I blasted the exons of the 21,777 annotated *C. pygargus* genes to the *Bos taurus* genome (Bovine Genome Sequencing and Analysis Consortium et al., 2009; Zimin et al., 2009; GCA_002263795.2), to the *C. capreolus* genome and to other published cervid genomes, namely: *Rangifer tarandus* (Li et al., 2017; PRJNA391754), *Odocoileus virginianus* (Seabury et al.,

2011; GCA_003697985.1, *Elaphurus davidianus* (Zhu et al., 2018; GCA_002443075.1), *Cervus elaphus* (Bana et al., 2018; GCA_002197005.1), *Odocoileus hemonius* (GCA_003697985.1), *Hydropotes inermus* (GCA_006459105.1), *Muntiacus muntjak (*GCA_006409035.1), *Muntiacus crinifrons* (GCA_006408485.1), *Muntiacus reevesi* (GCA_006408525.1) and *Cervus albirostris* (GCA_006408465.1) (Chen et al., 2019). During a second round of analyses I blasted the genes to five additional ruminant genomes, namely *Bison bison* (GCA_000754665.1), *Bison bonasus* (Wang et al., 2017), *Bos grunniens* (GCA_005887515.2), *Bubalus Bubalis* (GCA_003121395.1, Low et al., 2019), *Cervus canadensis* (Mizzi et al., 2017), *and Syncerus caffer* (Glanzmann et al. 2016). For each exon I selected the first hit only, giving preference to exons residing on the same contig or chromosome.

I used the getFastaFromBed tool from the bedtools (Edgar, 2004)version 2.19.1 to extract the blast hits from the reference genomes, and subsequently concatenated exons into whole genes (using the bash 'paste' command). I used Muscle (Edgar, 2004) version 3.8.31 for multiple alignments on the obtained gene sequences.

The credibility of the gene alignments was verified by visual inspection of a random subset of genes, as well as by using the concatenated alignments to generate a maximum likelihood species tree with 100 bootstrap replications using the software RaxML (Stamatakis, 2014) with *Bos taurus* as the outgroup, and with partitioning into first and second codon positions vs third codon positions.

dN/dS rates. I calculated gene specific dN/dS rates using PAML yn00 (Yang, 2007), opting for the yn method rather than the lwl85 method. For each gene, a species was excluded from the analysis if it contained a stop codon or 50 percent or more missing data points. I excluded genes of which the lengths were not multiples of 3.

Codeml branch site tests. I used PAML's CodeML to test for evidence of positive selection by comparing for each gene the performance of two branch-site models: model A ($\omega 0 < 1$, $\omega 1 = 1$, $\omega 2 > 1$), with the corresponding null model ($\omega 0 < 1$, $\omega 1 = 1$, $\omega 2 = 1$), by setting model = 2 and Nssites = 2 for both models, and fix_omega = 1 and omega = 1 when running the null model. Significance was evaluated using the chisq. test implemented in R, applying a Bonferroni correction for multiple testing.

I used three different foreground branches: *C. capreolus*, *C. pygargus* and the genus *Capreolus* (i.e. *C. capreolus* + *C. pygargus*).

To facilitate interpretation of outcomes and compare the findings to findings for other species pairs, in a second round of analyses I used additional foreground branches, namely: *B. bison, B. bonasus*, genus Bison, *C. elaphus, C. canadensis, C. albirostris*, genus Cervus, *B. bubalis*, *S. caffer*, subtribe *Bubalina*, *O. hemionus*, *O. virgianus*, and genus *Odocoileus*.

Accelerated dN/dS rates tests. PAML's codeML tests for the presence of positively selected codons within genes. It does so by comparing the likelihood of the null model that all codons within the gene are either evolving neutrally (i.e. dN/dS = 1) or under puryfing selection (i.e. gene wide dN/dS < 1) against the likelihood of the alternative model that in addition some codons are under diversifying selection (dN/dS > 1). Another way to search for positive selection is a relative rate test (Sarich and Wilson, 1973) which compares lineage specific gene specific dNdS rates to the background dNdS rates (i.e. gene specific dNdS rates in other lineages). If for a particular lineage multiple codons within a gene are under positive selection, the proportion of non-synonymous mutations in this lineage will be accelerated compared to other lineages.

I searched for genes with accelerated dNdS rates in *C. capreolus*, *C. pygargus*, and the *Capreolus* genus (i.e. *C. capreolus* + *C. pygargus*) using an ingroup-outgroup approach, which can be denoted in newick format as ((AB),C). For each of the three pairwise comparisons (A vs B, A vs C, and B vs C), I calculated the number of nucleotide and the number of amino acid differences. To search for accelerated dN/dS rates in species A, I contrasted the sums of the AB- and AC-scores to the BC-scores using Fisher exact tests.

Evidence for accelerated selection in *C. capreolus*, *C. pygargus* and the *Capreolus* genus was assessed using respectively the following ingroup-outgroup models: *((C. capreolus, C. pygargus), H. inermus), ((C. pygargus, C. capreolus), H. inermus)* and *((C. pygargus, H. inermus), R.tarandus)*.

GO enrichment analysis. GO enrichment analysis was executed using the R package systemPipeR (Backman and Girke, 2016). I downloaded the human GO

annotation file (i.e. goa.human.gaf) from the gene ontology consortium website (i.e. http://current.geneontology.org/products/pages/downloads.html) and used this dataset to create a catDB (using the function 'makeCATdb'). I chose the human GO annotation file because some genes of interests were missing from the cow GO annotation file. I executed GO enrichment tests using the functions GOHyperGALL and GOHyperGALL_Subset (GO slim analysis). Hugo gene ID's were converted to Swisprot gene ID's using the R package BiomaRt (Durinck et al., 2009).

Results

Genome wide heterozygosity. Genome wide heterozygosity estimates for *C. capreolus* and *C. pygargus* ranged between respectively 0.14-0.156% and 0.297-0.324%, depending on filter settings on mapping approach (Table A4.1, Fig. 4.1A-C,G).

Downsampling the *C. pygargus* dataset to the same read depth as the *C. capreolus* dataset lowered the He estimate of *C. pygargus* from 0.32% to 0.297% (Table A4.1). This outcome is suggestive of a false negative rate of heterozygous sites within *C. capreolus* (compared to C. *pygargus*) of 1 – (0.297/0.320) = 7.2%. Therefore, levelling the read depth of *C. capreolus* with the read depth for *C. pygargus* would potentially increase its He estimate from 0.143% to 0.154%.

Crossmapping sequencing reads to the reference genome of the sister species, returned heterozygosity estimates of 0.156% for *C. capreolus* and 0.324% for *C. pygargus* (Table A4.2).

Runs of homozygosity. Mean and median spacing between heterozygous sites was respectively 321 and 151 bp for *C. pygargus* and 761 and 240 bp for *C. capreolus* (Fig 4.1B). The two ROH-analysis methods produced consistent results (i.e. highlighted the same regions, Fig A4.6A-B). Using the He-spacing approach, we observed 23 genomic regions within the *C. pygargus* genome with low density of heterozygous sites, varying in length from 97.8 kb to 10.8 Mb, with mean spacing between heterozygous sites varying between 951.5 bp to 24.6 kb, and with a maximum

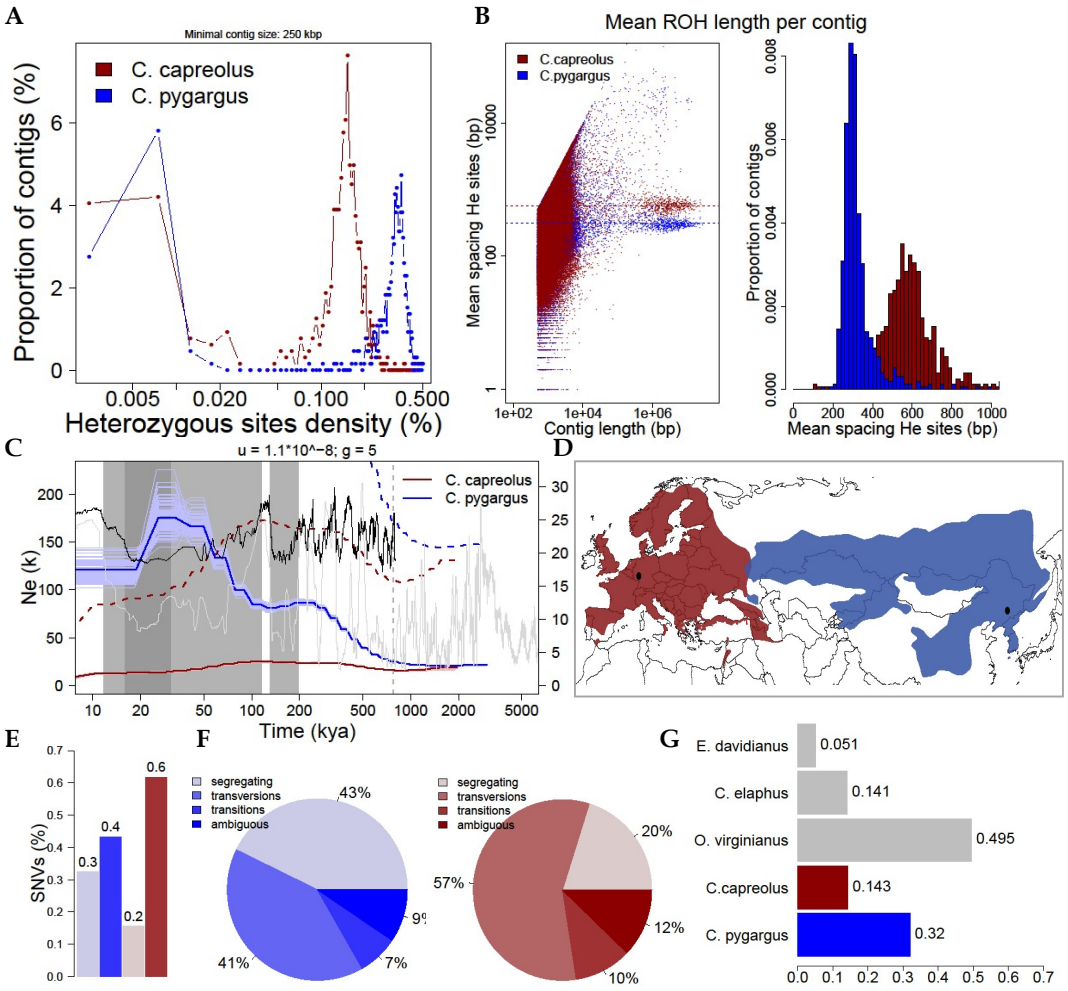


Fig. 4.1. Genetic diversity, genetic divergence and demographic history. *Red: European roe deer (C. capreolus), blue: Siberian roe deer (C. pygargus).* ***A****. Proportion of heterozygous sites per contig.* ***B****. Mean spacing between heterozygous sites. Left: all contigs. Right: contigs longer than 100kb.* ***C****. Historical Ne estimates inferred by PSMC analyses (Li & Durbin, 2011). Dashed coloured lines are relative to y-axis on the righthand size of the plot, full lines are relative to y-axis on the left hand side. Lightgrey area: Last Glacial Period (11.7-115kya) and Penultimate Glacial Period (130-194kya). Darkgrey area: Last Glacial Maximum (16.3-31kya).Dashed vertical line: Brunhes–Matuyama paleomagnetic reversal.Light grey line: magnetic susceptibility (/100), Lingtai Loess data (Sun et al. 2010). Black line: atmosphoric CO2/ppm/10, EPICA Dome C Ice Core 800kyr carbon dioxide data (Luthi et al, 2008).* ***D****.Geographic distribution of C. capreolus and C. pygargus. Data from IUCN website. Black dots indicate origins of samples from which whole genome sequences were obtained.* ***E****. Barplot of single nucleotide variations (SNVs) in C. capreolus compared to C. pygargus (left), and conversely C. pygargus compared to C. capreolus (right). Grey: segretating/heterozygous sites, colour: fixed sites.* ***F****. Piecharts of compostion of SNVs for C. capreolus compared to C. pygargus (left), and conversely C. pygargus compared to C. capreolus (right).* ***G****. Genome wide heterozygosity (percentage observed number of heterozygous sites of the total length of the genome assembly) of Capreolus species compared to other cervids. The estimate for E. davidianus was obtained from Zhu et al. 2018.*

spacing of 269.9 kb. Within the *C. capreolus* genome, I observed 101 genomic regions with low density of heterozygous sites, varying in length from 56.0 kb to 3.8 Mb, with mean spacing between heterozygous sites varying between 822.3 bp to 12.2 kb, and with a maximum spacing of 230.5 kb. Genomic regions with low density of heterozygous sites in *C. pygargus* did not overlap with those of *C. capreolus* (Fig. A4.2). The two species had similar variance in window He estimates (Fig. A4.6E), and the presence of a higher number of short ROHs in *C. capreolus* appeared consistent with the lower genome wide mean heterozygosity (Fig. A4.6E-F). When ignoring ROHs with lengths below 500Kb, the F_{ROH} estimates obtained for both species were close to zero (Fig A4.6D).

PSMC analyses. The output of the PSMC analysis indicated that throughout the separate histories of both sister species the historic effective population sizes (Ne) of *C. capreolus* has been consistently lower than the Ne of *C. pygargus* (Fig. 4.1, A4.3). The Ne of *C. capreolus* has remained roughly similar to the Ne of the ancestral population (i.e. ~20,000 individuals), whereas the Ne of *C. pygargus* has increased over time. Assuming a mutation rate of $1.1 \cdot 10^{-8}$ mutations per site per generation and a generation time of 5 years, *C. pygargus* Ne reached a maximum of ~175,000 individuals during the Last Glacial Maximum (i.e. 16-31 kya) (Fig. 4.2). Based on the same settings, the Ne estimates of the two species started to diverge around 1.5-1.6 Mya (Fig 4.2), suggesting *C. pygargus* and *C. capreolus* split before or at that time.

Higher mutation rates lead to more recent estimates of TMCRA, whereas longer generation times lead to less recent estimates of TMRCA (Fig A4.2). A rate of $2.5 \cdot 10^{-8}$ mutations per site per generation suggested a lower limit of the TMRCA of both sister species of 0.6 Mya, whereas a rate of $0.5 \cdot 10^{-8}$ mutations per site per generation suggested a lower limit of 3 Mya (Fig A4.3). A generation time of 3 years suggested a TMCRA lower limit of 0.7 Mya, whereas a generation time of 6 years suggested a TMRCA lower limit of 1.5 Mya (Fig A4.3).

Genome wide genetic divergence. Crossmapping *C. pygargus* sequencing reads to the *C. capreolus* reference genome yielded a pairwise sequence dissimilarity estimate of 0.60% (Table A4.2, Fig 4.1E-F). Crossmapping *C. capreolus* sequencing reads to the *C. pygargus* reference genome yielded a pairwise sequence dissimilarity

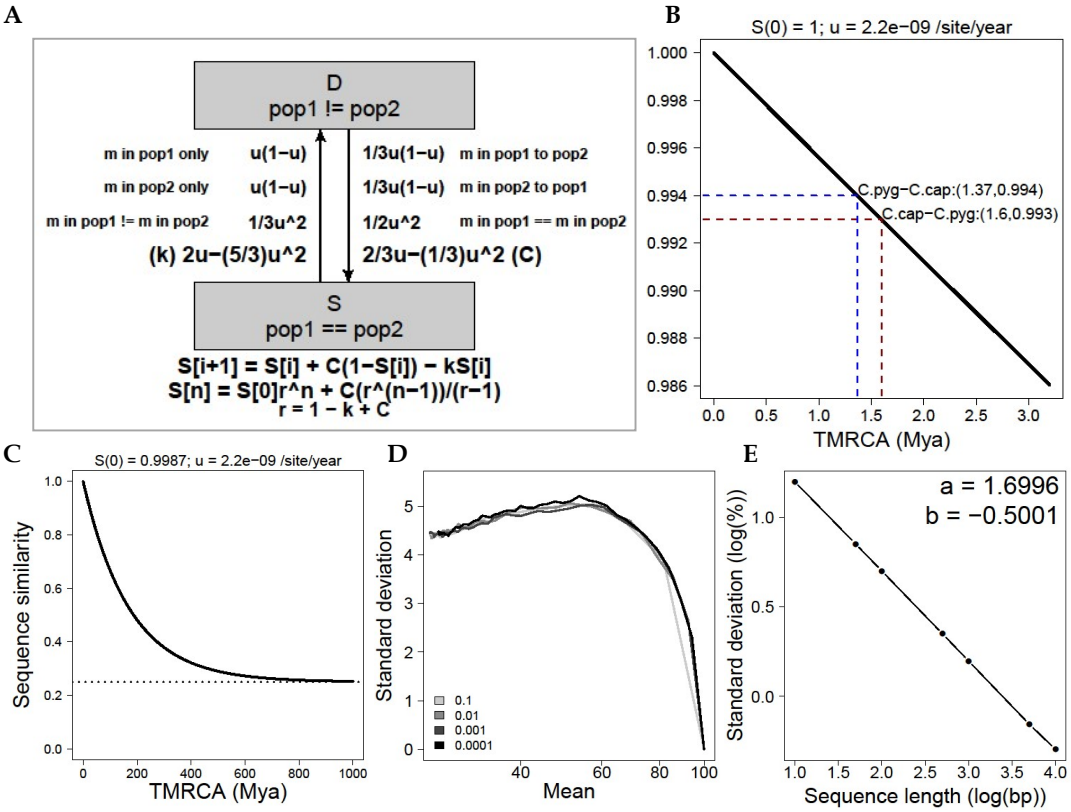


Fig. 4.2. Conceptual visualisation, simulation results and roe deer TMRCA estimate of the random walk Markov chain model. A. *Conceptual visualisation of the random walk Markov chain model. 'm' denotes a mutation event, 'u' denotes mutation probability per site per year or per generation,'i' and 'n' denote a single year or generation, 'pop1' and 'pop2' denote sister taxa which split at n=0 from ancestral population, 'D' denotes the sequence dissimilarity probability, 'S' denotes similarity similarity probability, 'S[0]' denotes sequence similarity probability directly after the vicariance event, and after fixation/loss of standing variation.* ***B.*** *Upper TMCRA estimate of C. pygargus and C. capreolus. C.pyg-C.cap: estimate derived from mapping C. pygargus reads to C. capreolus genome. C. cap-C.pyg: estimate derived from mapping C. capreolus reads to C. pygargus genome. Sequence similarity estimates (0.993-0.994)are based on analyses presented in 4.1E-F. S[0] is set to 1, resulting in upper TMRCA estimates.* ***C****. Sequence similarity decay predicted by the random walk Markov chain model. Sequence similarity converges as expected to 0.25, which is the sequence similarity of two unrelated DNA-sequences.* ***D.*** *Simulated confidence interval of sequence similarity estimates given a sequence of 100 bp. Sequence similarity estimates are generated with the recursive formula S[i+1] = S[i] + C*(1 – S[i]) – kS[i], in which C and k contain a stochastic element (i.e. occurence of mutation event). Shown are the mean and standard deviation obtained from 10,000 simulations with a 100 bp sequence, for a range of mutation rates (i.e. 0.1, 0.01, 0.001, and 0.0001).* ***E****. The relation between simulated sequence length and the maximum standard deviation of 10,000 simulated sequence similarity estimates. Simulated confidence interval of sequence similarity estimates given a sequence of 100 bp. Sequence similarity estimates are generated with the recursive formula S[i+1] = S[i] + C*(1 – S[i]) – kS[i], in which C and k contain a stochastic element (i.e. occurence of mutation event).*

estimate of 0.70% (Table A4.2, Fig 4.1E-F). Depending on the approach used, pairwise sequence similarity between *C. capreolus* and *C. pygargus* therefore equals either 99.3% or 99.4%.

Blasting *C. pygargus* genes to the *C. capreolus* genome resulted in pairwise alignments of in total 31,692,647 non-missing data points (sites with sequence information in both species), of which 170,596 sites were dissimilar, resulting in a exome dissimilarity score of 0.54%. This score is therefore approximately 10 percent lower than the genome wide dissimilarity score of 0.6% obtained from blasting *C. pygargus* reads to the *C. capreolus* genome.

Given the *C. capreolus* assembly measures 2,051,852,399 bp (Table A4.2), the exome made up (31,692,647/2,051,852,399*100=) 1.54% of the assembly. In contrast, exomic single nucleotide variations made up (170,596/(2,051,852,399*0.006)*100=) 1.39% of the genome wide number of single nucleotide variations.

Performance of split time estimation model. In line with expectations, the random walk MC model (Fig 4.2A) predicts a long term equilibrium neutral sequence similarity of 25% (Fig 4.2C). My similations indicate that the standard deviation from the expected similarity is independent of the mutation rate (Fig. 4.2D) and dependent on the length of the sequence (Fig. 4.2E) as well as the mean similarity. The maximum standard deviation is observed for a mean similarity of 50%, and is approximately described by the function: log(sd) = 1.7 – 0.5*log10(sequence_length) (Fig 4.2E). Therefore, given a genomic sequence of >1 Gb, the maximum standard deviation is 0.0016%, amounting to a very narrow 95% confidence interval of 0.0032%. This confidence interval is so narrow that I excluded it from the output plots. However, other factors do cause considerable uncertaincy in the estimation of divergence time, most notably mutation rate estimate error margins and sequence dissimilarity estimate error margins (Fig A4.5).

I validated the model by comparing the fit between expected (i.e. published) divergence time estimates for great ape species pairs with divergence time estimates outputted by the random walk MC model. Different outcomes were observed when using yearly mutation rates (u_y) versus generation specific mutation

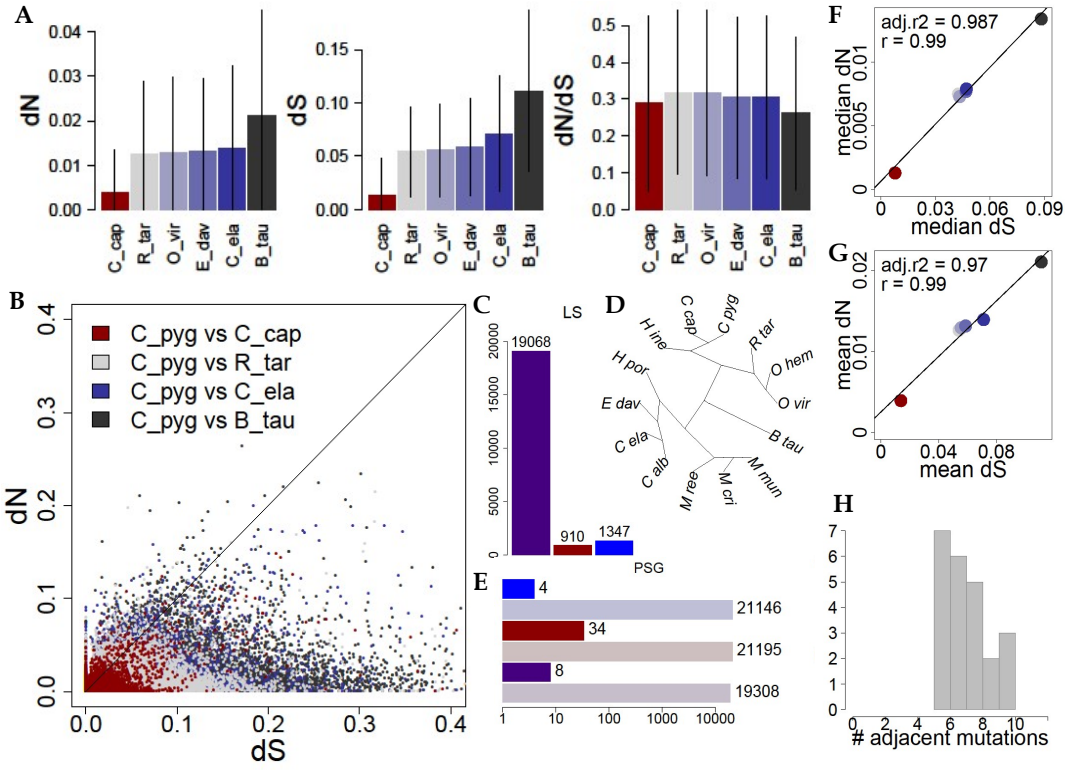


Fig 4.3. dN/dS analyses*. C_cap = C. capreolus (western roe deer), R_tar = R. tarandus (reindeer), O_vir = O. virginianus (white tailed deer), E_dav = E. davidianus (Pere David's deer), C_ela = C. elaphus (red deer), B_tau = B. taurus (cattle), H_ine = H. inermus (water deer), O_hem = O. hemonius (mule deer), M_mun = M. muntjak (common muntjac), M_cri = M. crinifrons (black muntjac), M_ree = M. reevesi (Reeves's muntjac), C_alb = C. albirostris (Thorold's deer).* ***A****. Barplots showing dN and dS values, calculated using PAML's yn00, for pairwise comparisons between C. pygargus and 5 other cervid species and cattle, for up to 14,512 genes. Bar heights indicate mean gene specific values. Error bars indicate standard deviation.****B****. Scatterplot of dN and dS values, calculated using PAML's yn00, for pairwise comparisons between C. pygargus and 3 other cervid species and cattle.* ***C****. Lineage sorting (LS) in Capreolus and Hydropotes depicted by frequency histogram of gene specific phylogenies for 21.325 genes, with R_tar as outgroup. Purple: ((C_cap,C_pyg),H_ine), red: ((C_cap,H_ine),C_pyg)), ((C_pyg, H_ine),C_cap).* ***D****. RaxML phylogeny based on full exomes with 100% bootstrap support at all nodes. I generated this phylogenetic tree to verify the gene alignments.* ***E****. Barplots of number of genes marked by codeml as neutrally evolving or positively selected genes (PSG). Light colour: neutral genes. Purple, red, and blue: PSG's for respectively genus, C. capreolus and C. pygargus as foreground lineages.* ***F.-G****. For all pairwise species comparisons, median dN and median dS values (as well as mean dN and dS values) are highly correlated, explaining the uniform dN/dS estimates across species, independent of TMRCA. Shown are adjusted squared explained variance and Spearman's correlation coefficient. Colour coding as in A.* ***H****. Barplot showing frequency of number of adjacent mutations in mutation clusters in genes marked by codeml branch site tests as outlier genes. Not counted are mutation clusters with gaps between the mutations.*

rates (u_g) (Fig A4.4). Given the 1.23% sequence dissimilarity reported for human and chimp genomes (Varki and Altheide, 2005), $u_y = 0.22 \cdot 10^{-8}$ results in a TMRCA estimate of 2.73 My and $u_g = 2.5 \cdot 10^{-8}$ results in a TMRCA estimate of 4.8 My (Fig. A4.4). Given the 0.6% sequence dissimilarity reported for bonobo and chimp genomes (Prüfer et al., 2012), $u_y = 0.22 \cdot 10^{-8}$ results in a TMRCA estimate of 1.26 My and $u_g = 2.5 \cdot 10^{-8}$ results in a TMRCA estimate of 2.2 My (Fig. A4.1). Given the 0.32% sequence dissimilarity reported for Sumatran and Bornean orangutans (Locke et al., 2011; Prado-Martinez et al., 2013), $u_y = 0.22 \cdot 10^{-8}$ results in a TMRCA estimate of 0.62 My and $u_g = 2.5 \cdot 10^{-8}$ results in a TMRCA estimate of 1.0 My (Fig. A4.4).

As for relatively similar species (i.e. sequence similarity > 99%) the random walk Markov chain model returned more faithful estimates using yearly rather than generation specific mutation rates (see results above), I decided to calculate the TMCRA of the *Capreolus* sister species using a yearly mutation rate (i.e.: $u_y = 0.22 \cdot 10^{-8}$).

Divergence time estimation. Uncertaincy in the estimate of the time to most recent common ancestor (TMRCA) of *C. capreolus* and *C. pygargus* arises from uncertaincy of three input variables: the estimate for sequence similarity shortly after the vicariance event (i.e. S(0) = 99.75 – 100%), the present day sequence similarity (i.e. S(n) = 99.3 – 99.4%), and the mutation rate. Assuming a yearly mutation rate of $u_y = 0.22 \cdot 10^{-8}$, the TMRCA of *C. capreolus* and *C. pygargus* ranges from 0.7-1.6 Mya, depending on combinations of S(0) and S(n) (Fig. 4.2B, Fig. A4.5).

Exome species tree. A species tree based on full exomes (with partitioning in first and second codon vs third codon positions) confirmed established relationships between cervid species, with *Capreolus* grouping together with Hydropotes in the New World Deer clade (Capreolinae) (Fig 4.3D, Fig A4.6). The *Capreolus*/*Hydropotes* clade contained the highest branch lengths, with *C. capreolus* having a higher branch length than *C. pygargus* (i.e. 0.0032 vs 0.0023) (Fig A4.6). Out of 21,325 gene trees with data for all three species, 86.7% (19,068 gene trees) corresponded to the species tree (i.e. ((*C.capreolus,C.pygargus*),*H.inermus*)), whereas 4.2% (910 gene

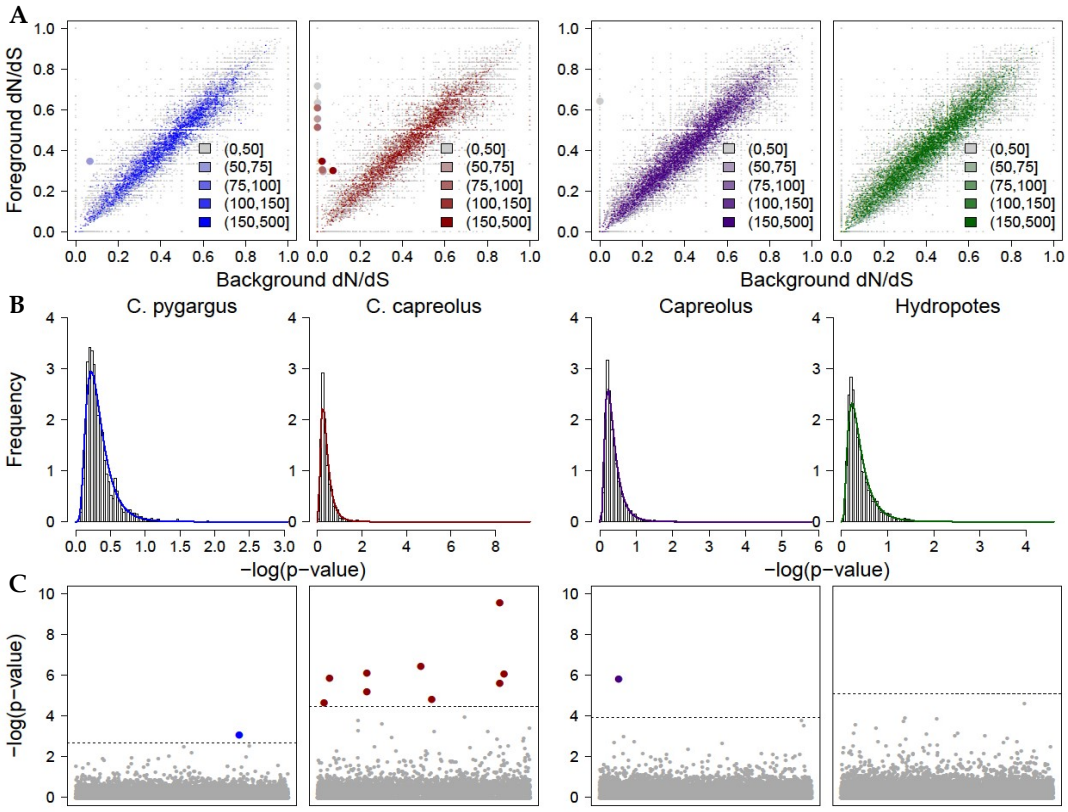


Fig 4.4. Accelerated dN/dS rates*. Gene specific proxies of dN/dS rates of foreground branches contrasted to gene specific proxies of dN/dS rates of background branches. Investigated foreground branches are C. pygargus (blue), C. capreolus (red), Capreolus genus (purple), and Hydropotes genus (green). All results are based on comparisons between three species, of which one species is defined as an outgroup species, as denoted in Newick format by ((A,B),C). Evidence for accelerated selection in C. capreolus, C. pygargus and the Capreolus and Hydropotes genera was assessed using respectively the following ingroup-outgroup models: ((C. capreolus, C. pygargus), H. inermus); ((C. pygargus, C. capreolus), H. inermus); ((C. pygargus, H. inermus), R.tarandus); and ((H. inermus, C. pygargus), R.tarandus). (**A**).For each of the three pairwise comparisons (A vs B, A vs C, and B vs C), I calculated proxies of dNdS ratios by counting the number of nucleotide and the number of amino acid differences. For each gene I summed the AB- and AC-scores (foreground dN/dS proxy), and contrasted these sums to the BC-scores (background dN/dS proxy). Colour coding indicates the sum of the observed nucleotide differences per gene. Inflated dots indicate genes marked as outliers (see 4.4C). (**B**). Gene specific AB/AC-scores were contrasted to BC-scores in a 2x2 contingency table, on which I subsequently executed Fisher exact tests. I found that the negative log of the Fisher exact p-values fits a lognormal distribution. Log mean and log standard deviations were respectively -1.22 and 0.53 for C. pygargus as foreground branch, -1.08 and 0.6 for C. capreolus as foreground branch, -1.17 and 0.58 for Capreolus genus as foreground branch, and -1.10 and 0.63 for Hydropotes as foreground branch. (**C**). I defined observed scores as outliers if they exceeded the 1-0.05/ngenes quantile threshold, with ngenes equalling 21777, and the threshold depending on the log standard deviation of the observed distribution (see 4.4B).*

trees) favoured *C. pygargus* as outgroup (i.e. ((*C_capreolus*,*H.inermus*),*C.pygargus*))), and 6.2% (1.347 gene trees) favoured *C. capreolus* as outgroup (i.e. ((*C_pygargus, H.inermus*),*C.capreolus*)) (Fig. 4.3C).

dN/dS rates. Mean and median dN and dS values for pairwise comparisons between *C. pygargus* and other cervids were strongly correlated (Fig. 4.3E-F) and dependent on TMRCA, with the pairwise comparison between *C. pygargus* and *C. capreolus* returning the lowest dN and dS values, and the pairwise comparison between *C. pygargus* and *B. taurus* returning the highest dN and dS values (Fig 4.3B,E-F). Due to the strong correlation between dN and dS values, dN/dS values were independent of the TMRCA. Mean and median dN/dS values for pairwise comparisons between *C. pygargus* and other cervids were relatively constant, ranging respectively between 0.26 and 0.32 (Fig 4.3) and between 0.14 and 0.16 (Fig A4.7), the mean values being roughly consistent with the expected long term equilibrium of 0.313 for genes under purifying selection, as inferred from simulations and modelling approaches (Mugal et al., 2014).

Codeml branch site tests. Codeml branchsite tests with the genus *Capreolus* as foreground branch (i.e. *C. capreolus* and *C. pygargus* combined) returned 18 out of 19318 genes with p-values below the Bonferroni threshold (Fig 4.3E,H; Table A4.3A). Visually examination of the gene alignments and the BEB-scores revealed that at least 10 genes were false positives due to either misalignments, missing data or paralog comparisons, leaving 8 genes (0.04%) with at least 1 or more lineage specific amino acid mutations with a BEB-score of 0.5 or higher for class2a or class2b (Table A4.3B-C; Fig A4.8). The 8 potentially positively selected genes were ARHGAP33, NLK, PAXBP1, MDN1, KLHL29, BOLA, ZCCHC18 and one undetermined loci.

Codeml branchsite tests with the species *C. capreolus* as foreground branch returned 70 out of 21,231 genes with p-values below the Bonferroni threshold (Table A4.4A). Visually examination of the gene alignments and the BEB-scores revealed that at least 39 genes were false positives due to either misalignments, missing data or paralog comparisons, leaving 34 genes (0.16%) with at least 1 or more lineage specific amino acid mutations with a BEB-score of 0.5 or higher for

class2a or class2b (Table A4.5B-C; Fig. 4.3E; Fig A4.9). Of those 34 genes, 25 genes were characterized by clusters of two or more adjacent amino acid mutations (i.e. >6 adjacent nucleotide mutations, Fig 4.5H), rather than by single mutations spread throughout the gene. Examples of these mutation cluster are depected in Fig A4.12. For 9 genes, these clusters of adjacent amino acid mutations were encoded for by DNA-sequences of 7 bp or longer which occurred once or multiple times elsewhere in the gene.

Codeml branchsite tests with the species *C. pygargus* as foreground branch returned 10 of 21152 genes with p-values exceeding the Bonferroni threshold (Table A4.6). Visually examination of the gene alignments and the BEB-scores revealed that at least 6 genes were false positives due to either misalignments, missing data or paralog comparisons, leaving 4 genes (0.02%) with at least 1 or more lineage specific amino acid mutations with a BEB-score of 0.5 or higher for class2a or class2b (Table A4.6B-C; Fig 4.3E; Fig A4.10). The 4 potentially positively selected genes were MAP1A, MUC2, NAP1L1, ZADH2. The gene MUC2 contained 47 amino acid mutations unique to *C. pygargus* (within the 14 species dataset), of which 3 adjacent codons, coded for by the 9-bp DNA-sequence 'CCACAACCA', which occurs at four other locations within this gene. The gene ZADH2 contained 5 amino acid mutations characteristic for *C. pygargus*, of which 3 adjacent, partly coded for by the 6-bp DNA-sequence GATGCA, which occurs at two other locations within this gene.

The outlier genes for *C. pygargus* and *C. capreolus* were not located in genomic regions with low density of heterozygous sites in the respective genomes.

Accelerated dN/dS rates. In line with expectations, gene specific foreground dN/dS rates generally correlated with gene specific background dN/dS rates (Fig 4.4A). The negative log of obtained Fisher exact test p-values (derived from comparing background dN/dS rates to foreground dN/dS rates) fitted a lognormal distribution (Fig 4.4B). I defined observed p-values as outliers if they exceeded the quantile threshold with a Bonferroni corrected p-value of 1-0.05/ngenes, with ngenes equalling 21777, and with the quantile threshold depending on the log standard deviation. I observed one outlier gene for *C. pygargus*, nine for *C. capreolus*, one for the *Capreolus* genus, and zero for the Hydropotes genus (Fig 4.4C).

Visual examination of the gene alignments revealed that two outlier genes were false positives due to paralog comparisons (Table A4.7). CodeML BEB-scores confirmed the abundant presence of codons with high class2a and class2b probabilities, but none of them were significant (Table A4.8, Fig A4.9), and the CodeML chi-squared p-values for these genes were highly insignificant (Table A4.7). The genes with accelerated dN/dS values were TMCC1 for the genus, DAGLB for *C. pygargus*, and SF3B1, MFAP1, EEF2, SLC16A7, SCN2A, SCN3A and PCSK2 for *C. capreolus*.

The outlier genes for *C. pygargus* and *C. capreolus* were not located in genomic regions with low density of heterozygous sites in the respective genomes.

GO enrichment analysis. GOslim analyses returned zero significant results for any of the gene outlier subsets. GO analyses returned 3 BP, 3 CC, and 3 MF GO terms with a p-value below 0.05 (after type 1 error correction) for *C. capreolus* codeml outliers; 9 BP terms for *C. pygargus* codeml outliers; and 1 CC and 1 MF term for the genus *Capreolus* codeml outliers (Fig $4.13A).

GO enrichment analysis for genes with accelerated dN/dS rates in *C. capreolus* returned respectively 7 BP, 8 CC and 7 MF GO terms with a p-value below 0.05 (after type 1 error correction, Fig A4.13B). The accelerated dN/dS rate tests resulted in less than three outlier genes for *C. pygargus* and the genus *Capreolus*, and did not return significant gene enrichment scores.

The majority of enriched GO terms for *C. capreolus* were represented by the gene pair SCN2A (g18675.t1, Uniprot: Q99250) and SCN3A (g18676.t1, Uniprot: Q9NY46) (Fig. A4.13A-B). These genes are also known as sodium-voltage gated channel alpha subunit 2 and sodium voltage-gated channel alpha subunit 3, and are predominantly expressed in the brain. Neither gene contained amino acid substitutions spread throughout the gene rather than clusters of substitutions. Closer inspection of the alignments revealed that parts of the sequences were shared between both genes and were suggestive of misaligned sections.

The majority of the enriched GO terms (all but one) for *C. pygargus* were represented by the gene pair NAP1L1 (g10234.t1, Uniprot: P55209) and MAP1A (g01212.t1, Uniprot: P78559), respectively known as assembly protein 1 like 1 and

microtubule associated protein 1A, which did not contain clusters of mutations (Fig. A4.13B).

The enriched GO terms for the genus *Capreolus* were 'cytosol' and 'transcription factor binding', represented by respectively four and two genes (out of a total of seven) known genes (Fig. A4.13A).

Discussion

This study compares the newly generated high quality reference genome of the Siberian roe deer (*C. pygargus*) to genomes of other deer species, most particularly the lower quality genome of it's sister species, the European roe deer (*C. capreolus,* NCBI assembly GCA_000751575.1, Kropatsch et al., 2013).

Genetic diversity. The genome comparison demonstrates a strong difference in nuclear genetic diversity between the two roe deer species, a finding which deviates from expectations based on comparisons of mtDNA studies. Reported control region nucleotide diversity estimates are 0.75% and 0.94% for respectively southwestern Germany (i.e. central European lineage (Baker and Hoelzel, 2014) and northeastern China (Lee et al., 2016), which are the sampling locations of the two whole genome sequences. The interspecies difference in nuclear DNA genetic diversity reported in this thesis (i.e. 0.14% and 0.32% heterozygosity in respectively *C. capreolus* and *C. pygargus*) is therefore almost twice the magnitude of genetic difference seen for mtDNA.

Given the relatively limited size of the mitochondrial control region (<1kb) in comparison to whole nuclear genome sequences (>2Gb), genome wide estimates provide more reliable estimates of genetic diversity. The comparatively weak difference in mtDNA genetic diversity likely reflects a genomic sampling bias. The contig specific estimates of heterozygosity confirm the presence of variation in genetic diversity along genomes, with some *C. capreolus* contigs containing equal or even higher genetic diversity than some *C. pygargus* contigs (Fig 4.1A-C).

Read depth has been shown to affect genotype calling. Homozygous SNVs are reliably detected at 15x coverage, whereas reliable detection of heterozygous SNVs requires a minimum depth of 18-20x (Meynert et al. 2014). Although the average read depth of *C. capreolus* was above 20, read depth varies stochastically across

sites, and in theory inaccurate genotyping calling at sites with read depths below 20 might have caused an underestimate of *C. capreolus* heterozygosity. However, downsampling of the *C. pygargus* dataset to the same average read depth as *C. capreolus*, only marginally lowered the He estimate for *C. pygargus* (Table A4.1), indicating that the strong difference in genome wide heterozygosity between both species is not a data artifact.

Given the wide geographical distribution of both species (Fig 4.1D) and the presence of isolation by distance effects (Baker and Hoelzel, 2012), the observed difference in genetic diversity does not necessarily reflect the species as a whole. MtDNA studies on *C. capreolus* and *C. pygargus* indicate considerable variation in genetic diversity across populations. Estimates of nucleotide diversity of the control region range between 0.00-0.82% for *C. capreolus* (Table 1 in Wiehler and Tiedemann, 1998; Table 2 in Baker and Hoelzel, 2014) and between 0.28-1.26% for *C. pygargus* (Table 2 in Lee et al., 2016). These figures indicate that although on average *C. pygargus* populations contain higher genetic diversity than *C. capreolus* populations, there is also considerable overlap.

Although the genome wide heterozygosity of *C. capreolus* is half the genome wide heterozygosity of *C. pygargus*, it is not exceptionally low, as it falls firmly within the range reported for other mammal species (see Fig 4A in Cho et al., 2013; Fig 1C in Robinson et al., 2016; Table S3 in Brüniche-Olsen et al., 2018; Fig 4A in Beichman et al., 2019). Also, pronounced differences in genome wide heterozygosity between closely related sister taxa have reported previously for other species, including great apes (Fig 1B in Prado-Martinez et al., 2013) and the extant two bison species (Wang et al. 2017; Brüniche-Olsen et al. 2018).

Run of homozygosity. Runs of homozygosity (ROH) analyses were performed to access whether the difference in genetic diversity between *C. pygargus* and *C. capreolus* could reflect a difference in inbreeding levels. The *C. pygargus* sample was obtained from a deer farm, and hypothetically the relatively high genome wide heterozygosity could have resulted from outcrossing between *C. pygargus* individuals originating from different geographic regions. Alternatively, the *C. capreolus* individual could have been an inbred individual. This latter explanation appears however unlikely, because the sample was obtained from a wild-caught

individual from a non-isolated population, and furthermore because the observed level of genome wide heterozygosity (~0.14%) corresponds with estimates obtained for >100 *C. capreolus* individuals from ddRADseq data (Chapter 3 of this thesis).

The outcome of the the ROH analyses also do not support the hypothesis that the *C. capreolus* individual was inbred. Although the *C. capreolus* genome does contain a higher proportion of regions with low heterozygosity (<0.01%) than the *C. pygargus* genome (Fig. A4.6C-D), this is likely not the result of inbreeding. Unlike long ROHs, which are likely to be autozygous as a result of recent inbreeding, short ROHs can also be caused by other factors, including historical population demography (Bruniche-Olsen, 2018). The vast majority of the ROHs detected in the *C. capreolus* genome were below 1Mb, and all were below 3Mb (Fig. A4.6C-D). In fact, the longest observed ROH, measuring 2Mb, did not occur in the *C. capreolus* but in the *C. pygargus* sample (namely on contig 16145, Fig. A4.6A-B). For comparison, samples with known history of recent inbreeding, such as cattle, contain ROHs stretching over 30Mb (Fig. 1 in Purfield et al., 2012). For both *C. capreolus* and *C. pygargus*, F_{ROH} estimates inferred from runs of homozygosity longer than 0.5Mb (Fig. 3 in Purfield et al. 2012) were (near) zero (Fig. A4.6C-D).

The more likely explanation for the difference in F_{ROH} estimates observed between the two *Capreolus* species is the difference in genome wide heterozygosity. Stochastic variation across genome translates into occurence of short ROHs in *C. capreolus* but not in the more genetically diverse *C. pygargus* (Fig A4.6E).

Because F_{ROH} estimates are defined as the proportion of ROH within genomes and hence can not be negative, F_{ROH} estimates do unfortunately not provide means to exclude the possibility of outbreeding (which would be relevant for the *C. pygargus* sample). However, it could be argued that an outbred individual will likely not contain a ROH of 2Mb.

Demography. Strict neutrality predicts that $He = \theta/(1+\theta) \approx \theta$ and that $\theta = 4 \cdot Ne \cdot u_g$ (Kimura, 1968; Kimura & Ohta, 1971). Therefore, under strict neutrality , genome wide He estimates can be converted directly into estimates of effective population sizes (Ne). Assuming a mutation rate per generation of $1.1 \cdot 10^{-8}$, the observed heterozygosities of 0.14% and 0.32% correspond with Ne estimates of respectively

32,000 and 73,000 individuals. Differences in the genomic distribution of heterozygous sites within the genome (Fig 4.1A) can reflect selective sweeps and/or differences in historic Ne, of which the latter can be inferred using coalescent modelling. The expected number of 32,000 individuals for *C. capreolus* falls slightly above the historical range of Ne values inferred by coalescent modelling (i.e. 8,000 - 25,000 individuals in the past 1 My, Fig 4.1C). In contrast, the expected number of 73,000 individuals for *C. pygargus* falls within the historical range inferred by coalescent modelling (i.e. 25,000 - 175,000 individuals in the past 1 My, Fig 4.1C).

The near convergence of the demographic trajectories of *C. capreolus* and *C. pygargus* around 1.5-1.6Mya provides a TMRCA estimate (Fig. 4.2B) which is in accordance with estimates obtained with the random walk Markov chain model. An upper estimate of 1.6 Mya differs from estimates based on mtDNA-studies, which resulted in lower and upper boundaries of respectively 2 and 4 Mya (Douzery and Randi, 1997; Randi et al., 1998; Xiao et al., 2007). MtDNA studies have previously been suggested to overestimate TMCRA (Lister, 2004) and my findings seem to support this conclusion.

TMRCA estimates inferred by the random walk Markov chain model span a wide range (0.7-1.6 My), which is mostly due to uncertaincy of the present-day sequence dissimilarity estimate. The sequence dissimilarity estimates calculated in this study depend on the cross mapping approach. The approach in which *C. capreolus* sequencing reads are mapped to the *C. pygargus* genome returns a higher sequence dissimilarity estimate (0.70%, Table A4.2) than the approach in which *C. pygargus* sequencing reads are mapped to the *C. capreolus* genome (0.60%, Table A4.2). The explanation for the observed discrepancy can perhaps be found in the differences in fixation time of mutant alleles in *C. pygargus* and *C. capreolus* as a result of different effective population sizes. An alternative explanation is that the number of fixed SNVs in *C. capreolus* is overestimated (and the number of segregating sites underestimated) due to the relatively low read depth of the *C. capreolus* dataset (Meynert et al. 2014).

Prado-Martinez et al (2013) used a similar approach to calculate sequence dissimilarity estimates between great apes population and species, and arrived at estimates of 0.32% for Bornean and Sumatran orangutan and 0.35-0.37% for chimpanzees and bonobos (Table S5.2 in Prado-Martinez et al., 2013). Divergence

times of these species pairs are estimated at ~1Mya (Fig 2 in (Prado-Martinez et al., 2013), suggesting that roe deer species (0.6-0.7% sequence dissimilarity) divergence time is more ancient.

The PSMC analyses suggest that the onset of the last glacial period (LGP) coincided with a decrease of *C. capreolus* Ne and in contrast an increase of *C. pygargus* Ne (Fig 4.1C). Contrasting historical demographic trends for closely related sister taxa have previously been reported for other species pairs, including common minke whales vs Antarctic minke whales (Kishida, 2017), common vs Indo-pacific bottlenose dolphins (Vijay et al., 2018), Bornean and Sumatran orangutans (Mattle-Greminger et al., 2018), northern and southern white rhino's (Tunstall et al., 2018) and chimpanzee and bonobo populations (Prado-Martinez et al., 2013). It is not clear why the demographic trajectories of the roe deer sister species should have been so different, but one possibility could be the differential impact of glaciations in the two regions. For example, in parts of Mongolia glaciers apparently retreated during the last glacial maximum due to the dry climate, in contrast to the expanding glaciers in Europe (Batbaabtar et al., 2018). Ancient DNA from a region nearby (the Denisova cave) confirms the presence of *C. pygargus* 21-50 kya (Vorobieva et al., 2011).

If climatic factors explain the different trajectories of the roe deer sister species, similar trends can perhaps be observed for other Eurasian mammals. Similar to the findings for *Capreolus*, PSMC analyses on *S. scrofa* genomes suggest a decrease of Ne in European populations with the onset of the LGP, and a concurrent temporary increase in Ne of Asian populations (Frantz et al., 2015; Groenen et al., 2012; Li et al., 2013). (The similarities in PSMC trajectories of *C. capreolus* and European *S. scrofa* sequences reflect similarities in the shape of the genome wide distributions of heterozygosity, consisting of a major frequency peak and two low diversity satellite peaks at at 0.02% and 0.005% heterozygosity (Fig 4.1A in this thesis, Fig. 2 in Groenen et al., 2012).) In contrast, the onset of the LGP does not coincide with a decrease in Ne for neither wisent (Fig 1. in Gautier et al., 2016; Fig S1 in Wu et al., 2018) nor red deer (Fig A4.2E), and also does not coincide with an increase in Ne of temperate Central and East-Asian deer species (i.e. Chinese muntjac, forest musk deer, white lipped deer and Pere David's Deer (Chen et al., 2019).)

The difference in demographic trajectory between *C. capreolus* and other European ruminants (i.e. wisent and red deer) might result from differences in feeding strategy. Whereas red deer and wisent are classified as intermediate feeders, roe deer are extreme concentrate selectors (Hofmann, 1989). Although the fossil record does not indicate strong differences between range shifts of *C. capreolus* (Sommer et al., 2009) and range shifts of red deer (*C. elaphus*) (Sommer et al., 2008), the different feeding strategies might have been differently impacted by the changing vegetation structures within the shifting ranges.

Chen et al. (2019) argue that a world wide trend of population declines in ruminants in the last 100 kya does not reflect climatic changes, but instead an increase of human activities around the globe. The observed decrease in *C. pygargus* Ne, around 25 kya, coincides with the colonization of *H. sapiens* of north eastern Eurasia (ref). In contrast, the observed decrease of *C. capreolus* in Europe, which sets in around 80 kya (Fig 4.1C), predates the arrival of modern humans.

The example of human-wildlife interactions illustrates that changes in effective population sizes do not necessarily reflect climatic changes, but can also result from ecological changes (i.e. different species interactions). Geographically separated environments typically contain different species assemblages even if they are environmentally similar. Differences in species community structures between western and eastern Eurasia could be implicated in driving the disparity of Ne estimates of *C. capreolus* and *C. pygargus*, rather than climatic conditions. A third possibility is that demographic fluctuations through time reflect selective events, as the ultimate response of adaptation is population size growth or decrease.

Apart from the coincidence of population size changes with climatic events in the past 100 kya (Fig 4.2B, Chapter 3 of this thesis), I find in general little relation between climatic transitions and *Capreolus* Ne (Fig 4.2B). No demographic changes are for example observed during the Penultimate Glaciation Period (Fig 4.1, Fig A4.2). This might suggests that the co-occurence of demographic and climatic changes in the past 100 kya is coincidental and should not be overinterpreted. An alternative explanation is that PSMC analysis lacks the resolution to detect demographic oscillations on the time frame of glacial-interglacial cycles, especially in distant times. Simulation studies indicate that Ne estimates of PSMC analyses do not faithfully represent recent (<100 kya) sharp and short lived (<50 kya)

bottlenecks, but instead suggest a constant intermediate Ne value for the bottleneck and pre-bottleneck time period (Spence et al., 2018). Short lived bottlenecks are more faithfully detected with SMC++ analyses (Spence et al., 2018).

***Effect of data quality and reference genome on demographic reconstruction*.** Historic Ne-values estimated by PSMC analyses have previously been shown to be more affected by read depth (Nadachowska-Brzyska et al. 2016) than by genome assembly quality (Patton et al. 2019). Consistent with previously published findings (Fig. 2 and 3 in Nadachowska-Brzyska et al. 2016), downsampling of the *C. pygargus* dataset to an average depth of 21 (and afterwards excluding all sites with a depth below 7) caused a downward and leftward shift of the inferred demographic curve, but did not affect the overall shape of the curve (Fig. A4.3A). In this study the lowest accepted read depth was set to 7. More accurate demographic trajectories would possibly have been obtained with a higher threshold of 10 or higher, even if as a consequence more data points would have been sacrifized (Fig 2 in Nadachowska-Brzyska et al. 2016).

PSMC curves generated by mapping *C. capreolus* reads to the *C. pygargus* genome assembly resulted in a different demographic trajectory compared to the demographic trajectory obtained by mapping *C. capreolus* reads to the *C. capreolus* assembly (Fig A4.3B). Simulation studies indicate that genome assembly quality (i.e. scaffold lengths) does not not impact inference of population size history (Fig. 3 in Patton et al. 2019), which suggests that the different trajectories observed for *C. capreolus* do not result from the difference in assembly qualities between the *C. capreolus* genome and the *C. pygargus* genome. A difference was observed in the proportion of heteroyzygous sites inferred from mapping *C. capreolus* reads to the *C. pygargus* genome (0.156%) compared to the estimate obtained from mapping *C. capreolus reads* to the *C. capreolus* genome (i.e. 0.143%). This difference of nearly 10% could explain the different trajectories, but it is unfortunately not clear what caused the difference in He estimate, and neither why as a result Ne estimates would differ most strongly in most recent and most distant times (Fig A4.3B).

Exome evolution. Whatever caused the difference in Ne between the *Capreolus* sister species, one interpretation of the exome tree and the results of the genic

selection analyses is that the difference in Ne impacted genic evolution in the *Capreolus* sister species. The exome species tree indicates a higher branch length for *C. capreolus* than for *C. pygargus* (i.e. 0.0032 vs 0.0023, Fig A4.6). Similarly, the accelerated dN/dS tests indicate a higher proportion of non-synonymous changes within *C. capreolus* than for *C. pygargus* throughout the exome (as indicated by differences in the shape of the Fisher exact test p-value distribution, Fig 4.4B-C).

Whereas positive selection targets specific or subsets of genes only, demography affects the entire genome (Lewontin and Krakauer, 1973). Observed exome wide differences are therefore likely accounted for by demographic differences rather than by positive selection, in which case the elongated exome branch length of *C. capreolus* reflects relaxed purifying selection of nearly neutral (i.e. mildly deleterious) mutations in *C. capreolus* due to its relatively low historic population sizes (Kimura et al., 1963; Ohta, 1992; Ohta and Gillespie, 1996). Consistent with this explanation, the exome species tree (Fig. A4.6) predicts higher branch lengths (i.e. 0.0032 vs 0.0018) for *C. elaphus* (red deer, long term Ne <= 25000, Fig A4.3) than for the historically more abundant sister species *C. albirostris* (white lipped deer, long term Ne ≈ 100,000, Fig S31 in Chen et al., 2019. However, if the observed violation of a steady global molecular clock is indeed caused by differential fixation of deleterious mutations, this violation can be expected to be attenuated when branch lengths are calculated for third codon positions only. In contrast, the variation in substitution rates among branches observed for 3rd codon positions did not differ from whole codon substitution rates (Fig A4.6). Additional comparisons between sister species with contrasting Ne, as well as a more robust estimate of tree topology and statistical testing, would be needed to confirm the hypothetical relationship between exome branch length and effective population size.

Kimura (1962) estimated the fixation probability of a negatively selected allele as: exp(4*Ne*s*(1/Ne))-1)/(exp(4*Ne*s)-1). For Ne = 1000 and s = 0.0001, the fixation probability equals $3.2*10^{-4}$, whereas for a population with Ne = 100000, the fixation probability equals $1.7*10^{-20}$. This means that in a population with Ne = 1000, 1 out of 3000 nearly neutral mutations will fixate, whereas in a population of Ne = 100000, fixation of any mutation is highly unlikely.

It was found that the the exome-wide average ratio between the non-synonymous (replacement) and synonymous (silent) substitution rate (i.e. dN/dS or ω), was independent of divergence time, and equalled around 0.3 for all investigated species pair comparisons (Fig 4.3A). The observed absence of correlation between average dN/dS and TMRCA corresponds with previously published findings (Fig 1 in Nei, Suzuki and Nozawa, 2010). Based on fixation probability equations for selected and neutral alleles (Kimura, 1962), it can be mathematically shown that following the split of two sister populations/species, their average pairwise dN/dS ratio will converge within 20·Ne generations to an asymptotic value determined by the scaled selection coefficient, which is the product of the effective population size and the average selection coefficient acting on replacement sites (i.e. $\Upsilon = Ne \cdot s$, Mugal et al., 2014, Why time matters).

An average exome wide dN/dS value of 0.3, and an associated scaled selection coefficient of -1, is close to previously published estimates. Comparisons between giraffe, okapi, and cattle genomes generated average dN/dS estimates of 0.22 (Agaba et al. 2016). A meta-analysis on pairwise comparisons between humans and a range of other vertebrate species resulted in average dN/dS ratios below 0.3, converging to approximately 0.1 for species pairs with the deepest divergence times (Wolf et al. 2009). (The observed correlation between dN/dS and TMRCA might result from a correlation between TMCRA and effective population.)

Expected dN/dS-values depend on proportions of deleterious, neutral and adaptive mutations (defined as respectively prop(d), prop(n), and prop(a)), and can be calculated for various scenarios using fixation probability functions (Fig. A4.14, Kimura, 1962, Mugal et al. 2014). From comparison between the observed dN/dS values (which range between 0.1 and 0.3) and expected dN/dS-values, three conclusions can be drawn:

1. Observed dN/dS-values are not consistent with prop(s) ≥ 0.0025, indicating that the proportion of positively selected non-synonymous mutations is below 0.25%.
2. Observed dN/dS-values are not consistent with prop(n) ≥ 0.5, indicating the majority of neutral non-synonymous mutations are deleterious and under purifying selection.

3. Observed dN/dS values are not consistent with Y > -1, indicating that the magnitude of the selection coefficient acting on deleterious mutations is bigger than the inverse of the effective population size (i.e. |s| > 1/N).

The first two conclusions are consistent with the neutral theory, which holds that most non-synonymous mutations are deleterious (King and Jukes, 1969; Kimura and Ohta, 1971). The latter conclusion is consistent with the nearly neutral theory, which holds that deleterious mutations of which the magnitude selection coefficient is smaller than the inverse of the effective population size, are effectively neutral.

False positive rates of codeml selection scans. The genes marked by codeml as possibly having experienced episodic selection in either of the three roe deer lineages, could be divided into two groups: genes containing lineage specific single nucleotide substitutions spread throughout the gene, and genes containing clusters of directly adjacent nucleotide substitutions. The minority of genes belonging to the first category predominantly surfaces in the GO enrichment analyses (Fig A4.13).

The majority of codeml outlier genes, namely 34 out of 46 genes, belonged to the second category, Positive selection has been argued to be able to cause multiple amino acid substitutions in close proximity (Wagner, 2007; Zhou et al., 2008). However, two alternative explanations exist: data artifacts, and relaxation of purifying selection.

Data artifacts in codeml input datasets can originate from four potential sources, namely from genome sequencing errors (Mallick et al., 2009; Schneider el al. 2009), annotation errors, blasting errors and alignment errors (Fletcher and Yang, 2007; Jordan et al. 2012; Harrison et al. 2014). The codeml false positive rate due to blasting, annotation and alignment errors is presumably independent of the selected foreground branch, whereas in contrast false positive rates can be expected to depend on the quality of the genome assembly of the species selected as foreground branch. Mallick et al. (2009) concluded for example that the initially relative high number of inferred genes under positive selection in the chimpanzee lineage was caused by the relatively low quality of the chimpanzee genome sequence, and that the signal of selection for most outlier genes disappeared after generating a higher quality chimpanzee genome assembly. Schneider el al. (2009) found that genes with low coverage, annotation and alignment scores were

considerably more likely to be marked by codeml as outlier than genes with high scores. Fletcher and Yang (2010) found that the codeml false positive rate due to alignment errors depended on the alignment tool, and that the lowest false false positive rate is obtained when using the alignment tool PRANK. Alignment filters, such as SWAMP (Harrison et al. 2014) can mitigate false positive rates to a certain extent (Jordan et al., 2012).

Because the gene datasets in this study were generated with the alignment tool MUSCLE (rather than with the alignment tool PRANK), and because no alignment filter was applied, false positive rates are likely high. Furthermore, because the *C. capreolus* genome is of lower quality than the *C. pygargus* genome (e.g. 24x coverage vs 100x coverage), higher false positive rates could be expected for *C. capreolus* than for *C. pygargus*, which might explain the differences in number of codeml outlier genes found for both lineages. If this explanation is true, the observed clusters of adjacent substitutions are not adaptive gene modifications, but instead data artifacts, which originated from sequencing errors, and which do not truly exist in nature. This hypothesis is supported by the observation that some of the amino acid substitutions are highly unlikely according to Dayhoff matrices of amino acid transition probabilities. The hypothesis was investigated by generating counts of codeml outlier genes for a range of foreground lineages with various genome sequence qualities, estimated by average coverage. If a relationship exists between number of outlier genes and sequencing depth, this relationship is non-linear and confounded by other factors (Table A4.9; Fig. A4.15).

Positive selection vs relaxed purifying selection. Even if the observed clusters of adjacent substitutions are not data artifacts, these substitutions are not necessarily driven to fixation by positive selection. Substitutions can also occur due to relaxation of purifying selection (He et al., 2018). Both the codeML branchsite tests (Fig 4.3D,G) and accelerated dN/dS rate tests (Fig 4.4) identified more potentially positively selected genes (PSGs) in the *C. capreolus* lineage than in either the *C. pygargus* lineage or the genus *Capreolus* lineage. Branch site test with other ruminant species and genera as as foreground branch, consistently produced the same results: less outlier genes were found for genera than for species (Table A4.9).

Finding more lineage specific outlier genes in species than in genera seems to contradict expectations based on ecological and phenotypic comparisons. *C. capreolus* and *C. pygargus* species, for example, differ from other cervids, including the sister genus *Hydropotes*, more strongly than they differ from each other. Therefore most functional changes within the genome – within either genes or regulatory sequences – should be expected to have occurred before rather than after the split of *C. capreolus* and *C. pygargus*. It seems therefore reasonably to expect higher dN/dS rates (and hence a higher number of codeml outliers) if the genus lineage rather than for either of species lineages. (See for example observed lysozymes dN/dS rates between and within lineages of foregut fermenting and non-foregut fermenting primates (Messier and Stewart, 1997.)) In contrast, the 'relaxed purifying selection hypothesis' does not predict such a correlation between phenotypic and genomic divergence.

Substitution clusters can arise through single multinucleotide mutation (MNM) events (Schrider et al., 2011). Recently it has been reported that these MNMs cause codeml to incorrectly infer positive selection (Venkat et al., 2018), and I independently considered the same conclusion. Genes with high density of independent amino acid mutations are more likely to be driven by positive selection than by relaxed purifying selection, but this is not necessarily the case for genes which stand out because of a single mutation, such as a single multinucleotide mutation event. It is therefore not impossible that the majority of genes (34 out of 46) marked by codeml as having experienced positive selection could in fact be accounted for by relaxation of purifying selection.

Characterisation of MNMs in human genomes indicated that the vast majority of MNM's are 2-bp mutations, followed by 3-bp, 5-bp, 4-bp mutations, with a small minority representing 7-bp and 8-bp mutations (Fig 2b in Besenbacher et al., 2016). The mutation clusters in codeml outlier genes observed in this study generally ranged between 5 and 10 bp (Fig 4.3H), and the mutations in these clusters were all directly adjacent (i.e. no spacing in between). In several cases I found the nucleotide sequences coding for the observed clusters of amino acid substitutions to occur once or several times in other parts of the gene. In one extreme case, I found a 9 bp sequence responsible for 3 adjacent amino acid substitutions (CCACAACCA) to occur four times elsewhere in a gene of 7kb. This could suggest that the source of

the adjacent non synonymous substitutions were single events of translocations of sequence blocks of ~5-10 bp length, rather than accumulations of single nucleotide mutations. However, insertions of a sequence block often leads to a frame shift mutations, and this was not observed in the outlier genes. Furthermore, if relaxed purifying selection is responsible for the accumulation of the substitution clusters and for the relatively high number of codeml outlier genes in *C. capreolus*, a negative correlation is expected between estimates of genetic diversity for a certain foreground branch and the number of codeml outlier genes. No such relation was observed for the small ruminant dataset generated in this study (Table A. 4.9; Fig. A4.15).

The relation between Ne and genetic load. The 'relaxed purifying selection'-hypothesis holds that the fixation probability of slightly deleterious mutations is a function of effective population size (Ohta, 1992), and hence that genetic load is a function of effective population size (Ne). However, causality runs both ways, as population size (both effective and census) are a function of fitness and thus of genetic load.

In a non-changing environment and in the absence of genetic drift, the arrival of an adaptive mutant allele will lead to a population size increase, because the affected individual has the potential to produce more offspring than other members in the population. This population size increase will until all individuals in the population contain the adaptive allele, at which point all individuals have a higher reproductive rate than prior to the mutation event.

The same logic applies, conversely, to the effect of deleterious alleles. An individual which carries a mutant deleterious allele has a deterministically lower reproductive output than other members of the population. If through stochastic factors (i.e. genetic drift and/or genetic hitchhiking) this deleterious allele becomes fixated in the population, all individuals in the population have a lower reproductive rate than prior to the arrival of the deleterious allele.

Therefore, an alternative explanation for the observed differences between *C. capreolus* and *C. pygargus* (i.e. lower Ne and higher exome branch length for the *C. capreolus* lineage). Apart from being a data artifact or resulting from relaxed purifying selection, the inverse correlation between Ne and number of amino acid

substitutions could perhaps be caused by the concept of genetic load. This alternative explanation involves a hypothetical scenario in which shortly after the split of both incipient sister species, *C. capreolus* experiences a population bottleneck with stochastic fixation of deleterious mutations. Theoretically, the presence of these deleterious mutation could cause a permanent reduction in Ne, even when the cause of the population size bottleneck has disappeared.

This hypothesis, although admittingly highly speculative, could potentially explain the puzzling differences in effective population sizes observed between closely related and ecologically similar sister species, such as *C. capreolus* and *C. pygargus*. Given the possibility of back mutations, it appears questionable if the duration of the fitness effect reduction could have persisted throughout the life span of *C. capreolus*. In theory, a positive feedback loop consisting of relaxation of purifying selection, increase of genetic load and decrease of population size could dwindle a population towards extinction. In contrast, PSMC analyses suggest that *C. capreolus* Ne remained relatively constant. Furthermore, if *C. capreolus* individuals would indeed have had a lower fitness than *C. pygargus* individuals, *C. pygargus* individuals could be expected to replace *C. capreolus* individuals at the hybrid zone, leading to a gradual shift of range boundaries, leading to the eventual disappearance of *C. capreolus*. Simulation studies – for example with the software SLIM (Haller and Messer, 2019) – could serve to test these expectations, and more generally to investigate the interaction between Ne and genetic load.

Number of positively selected genes (PSGs). The proportions of genes marked by codeml branch site tests as outliers for the three foreground branches (i.e. 0.02% for *C. pygargus*, 0.16% for *C. capreolus* and 0.04% for *Capreolus* genus) are consistent with published estimates. For example, codeml branch site tests with as foreground branches Bornean and Sumatran orangutans, which split ~1 Mya and like *C. capreolus* and *C. pygargus* exhibit limited ecological differentiation, resulted in respectively 46 (0.14%) and 33 (0.10%) outlier genes out of 34,379 exonic sequences (Mattle-Greminger et al., 2018).

If ecological and phenotypic differentiation is driven by substitutions within genes, higher proportions of outlier genes were to be expected for the comparison between species with high niche differentiation. In reality, the reported number of

genes marked by codeml as having experience episodic positive selection in species with high niche differentiation do generally not differ from the findings for *C. capreolus* and *C. pygargus*. Transcriptome sequencing of red fox and arctic fox, which diverged ~3 Mya, revealed 4 genes (0.08%) marked as outliers by codeml branch site tests in red fox and another 8 (0.16%) in arctic fox, out of 4,937 genes in total (Kumar et al., 2015). Another study reported 10 (0.07%) and 18 (0.12%) genes out of 14,558 genes being marked by codeml as outliers in respectively humans and chimpanzees, with an additional 7 (0.06%) out of 10,980 genes being marked as outliers for the most recent common ancestor of humans and chimps (Kosiol et al., 2008). In a study on big cats, codeml identified 31 (0.24%) outlier genes in jaguar, 4 (0.03%) in lion, 3 (0.02%) in snow leopard, and 149 (1.13%) in tiger, out of 13,183 genes in total (Figueiró et al., 2017). This latter estimate is below the outcome of an earlier study, which reported 178 (2.40%) outlier genes out of 7,415 genes (Cho et al., 2013).

The overlap in number of codeml outlier genes between phenotypically conserved and phenotypically diverged species might indicate that phenotypic divergence is marginally driven by substitutions within genes, and predominantly by other genomic changes such as gene copy number variations (Rinker et al., 2019), mutations in regulatory sequences (Brawand et al., 2014; King and Wilson, 1975; Sackton et al., 2019), de novo gene evolution (Baalsrud et al., 2018) and gene silencing through genomic translocations (Hof et al., 2016).

The power of codeml branch-site tests has been shown to depend on multiple factors, and to be strongly positively correlated to gene sequence length, the proportion of codons under positive selection (defined as p2), the strength of positive selection, and weakly to the length of the foreground branch and the number of included sequences (Yang and Reis, 2010). Simulation analyses have shown that for an average sized gene (500 codons, Yang and Reis, 2010), a proportion of 10 percent positively selected codons (i.e. p=0.1), and a scaled selection coefficient of 2 (i.e. s = 2/Ne and dN/dS = 4), codeml power estimates range between 0.6 and 0.8 (Table 3), depending on the length of the foreground branch and the number of sequences (8 or 16) included in the analysis (Table 3 in Yang and Reis, 2010). For the *Capreolus* study species (Ne > 25000), a scaled selection coefficient of 2 roughly corresponds to a selection coefficient of 0.0001, with is

reasonably low. However, the proportion of codons under positive selection is likely lower than 10%, and for proportions below 5%, (i.e p2 < 0.5), codeml power estimates drop below 0.2 (Fig. 4c in Yang and Reis, 2010). Therefore, if the proportion of codons under positive selection within a positively selected gene is generally below 5%, the number of codeml outlier genes might be considerably lower than the true number of positively selected genes. Codeml false positive rates appear unrelated to sequence length, number of sequences and length of the foreground branch, and deviates, according to simulations, around 5% (Table 2 in Yang and Reis, 2010).

One limitation of branch-site models implemented in codeml is that they do not account for variation in synonynomous and non-synonymous substitution rates within foreground branches and/or within background branches (Murrell et al. 2015). This limitation has been put forward as explanation for the counterintuitive observation that inclusion of additional sequences can cause codons to be no longer marked as being putatively under selection (Murrell et al, 2012). It might also explain why the number of outlier genes inferred for genera is generally below the number of outlier genes inferred for species lineages (Table A4.9), although an alternative explanation is that genotyping errors are unlikely to cause the same data artifacts in different genome sequences. New selection scans tests have been developed which make use of improved underlying models which assume that substitution rates can vary between each lineage and between each site (Murrell et al. 2012; Murrell et al. 2015). These tests are claimed to have higher power than codeml (Murrell et al. 2015), and therefore to provide a more accurate picture (i.e. lower false negative rate) of the number of genes under positive selection.

Another assumption of codeml branch-site tests which is likely frequently violated, is the assumption that variation in non-synonymous substitution rates is negligible (Wisotsky et al. 2020). Relaxation of this assumption leads to lower false positive rates (Wisotsky et al. 2020), and thereby can furthermore increase the accuracy of the estimate of the number of genes under positive selection.

Recommendations for future selection scan studies. In conclusion, the results of the selection analyses are suggestive of positive selection events, but can also not exclude the possibility of relaxation of purifying selection and false positives due to

data artifacts. Additional comparative genomics studies of species diversifications are needed to make stronger inferences.

Future studies aiming to employ branch-site tests to assess the influence of natural selection on protein coding DNA during the formation of species and genera, should consider the following recommendations:

- Usages of high quality genome sequencing datasets minimizes false positive rates. Unreliable genotype calls should be filtered out by masking all sites within a genome with a coverage below 18-20 (Meynert et al. 2014). If site-specific read depth information is not available, genomes included in analyses should have a read depth well above 20, to ensure that sufficient read depth for the vast majority of sites.
- Each lineage/species should be represented by multiple samples/individuals. Including multiple samples per lineage provides a certain leverage to discriminate false positives (for example genotyping errors) from true codons under selection.
- Future comparative genomic studies should ideally not only encompass comparisons of coding sequence but also comparisons of regulatory sequences.
- To minimize false positives caused by alignment errors, genes should be aligned with the alignment tool PRANK (Fletcher and Yang, 2010; Jordan et al. 2012).
- Gene alignments should subsequently be filtered using alignment filtering tools, such as SWAMP (Harrison et al., 2014).
- Each gene marked by codeml as outlier, should be inspected visually, specifically the codons which are putatively under selection (as can be inferred from the codeml BEB-tables).
- Apart from the model implemented in PAML codeml, new models have been developed to detect PSGs. A range of models – including BUSTED for gene-wide selection, aBSREL for lineage-specific selection, MEME for site-specific episodic selection and FUBAR for site-specific pervasive selection – are implemented in the software HyPhy (Pond et al. 2005; Murrell et al., 2012; Wisotsky et al. 2020). These models are more sophisticated, and are thought

to give both lower false negative rates (Murrel et al. 2012) and lower false positive rates (Wisotsky et al. 2020).

- In addition to using multiple codon-based selection scans, it is also advisable to use other types of selection scans. Genes can stand out in various ways, ranging from the presence of individual lineage specific codons to genes with substitution clusters, and to the presence of genic regions with accelerated dN/dS rates or accelerated substitution rates in general (i.e. accelerated dN rates paired with accelerated dS rates). As these genes exhibit different signals, complementary selection scans are needed to identify all of them. In the case of accelerated substitution rates, relative rates tests such as the one presented in this thesis Chapter could proof useful.
- Finally, it should not be assumed that substitutions are driven by positive selection without considering the alternative explanation of relaxation of purifying selection (He et al., 2018). A tool specifically designed to discriminate relaxed purifying selection from increased positive selection is the software RELAX (Wertheim et al. 2015). The extent of relaxed purifying selection can also be estimated using the GERP-score (genomic evolutionary rate profiling, Cooper et al, 2005), implemented in the software GERP++ (Davydov et al., 2010). The GERP score quantifies the difference between the observed and the expected number of substitutions within a lineage. The expected number of substitutions are estimated based on a multi-species sequence alignment and a given phylogeny containing TMRCA estimates between aligned species. Because a GERP-score is effectively an estimate of the number of rejected substitutions, it quantifies the strength of past purifying selection, and hence can be used to assess the likelihood that substitutions within outlier genes are caused by relaxed purifying selection rather than by positives selection.

Conclusions

I estimate that *C. capreolus* and *C. pygargus* started to diverge at maximum 1.6Mya. Genome wide heterozygosity in *C. pygargus* is twice as high as genome wide heterozygosity in *C. capreolus.* PSMC analyses indicate that after the split, *C. pygargus* Ne gradually increased from 20k to a maximum of 170k, whereas *C.*

capreolus Ne has fluctuated around 15-20k. Correlations between climatic transitions and demographic changes inferred by PSMC are restricted to the last 100 ky. *C. capreolus* genes contain a higher proportion of both synonymous and non-synonymous substitutions compared to *C. pygargus* genes, which might reflect data artifacts or a combination of episodic positive selection and relaxation of purifying selection commonly associated with small population sizes.

Chapter 5
General discussion

In this thesis I have presented the outcomes of genetic analyses of several reindeer and roe deer datasets, using two types of data: single nucleotide polymorphism (SNP) data and whole genome sequencing data. Although my analyses also assessed the population structure, genetic diversity and demographic history of the study populations and study species, the focus was on selection analyses – the detection of genetic signals of selection.

In this last Chapter I will discuss how the findings presented in this thesis compare to expectations of the (nearly) neutral theory, which holds that most differences between populations and between species in protein-coding and regulatory DNA are caused by fixation of (nearly) neutral alleles rather than by fixation of adaptive alleles.

Overview of results presented in this thesis

In Chapter 2 I described a study in which I searched for shared signals of selection in two reindeer founder populations. These two populations were founded at the start of the 20th century, when whalers released two small herds of reindeer on geographically separated peninsula of the South Atlantic island South Georgia. Because the populations were founded in parallel in similar environments without the possibility of gene flow, they provided a suitable study system to overcome the complications associated with the detection of empirical evidence for natural selection in bottlenecked founder populations.

I harnessed the double digest restriction-site associated DNA sequencing (ddRADseq) protocol to generate an 80K SNP dataset of both founder populations as well as of their common Norwegian source population. I screened this dataset for signals of selection using four different selection scans: Bayescan, OutFLANK, PCadapt and a custom-built tool which I named Genome Wide Differentiation Scan (GWDS). Three SNPs were identified as outliers by two or more selection scans. Alignment to a reindeer reference genome indicated that two outlier SNPs were adjacent and 80 kB apart.

To evaluate the possibility of false positives, I performed forward-in-time simulations of frequencies of neutral and adaptive alleles in order to estimate the power and specificity of selection scans in the context of founder populations. These simulations indicated that loci under positive selection in non-communicating sister founder populations are most confidently detected by GWDS, and that SNPs marked as outliers by multiple selection scans are most likely true loci under selection. In summary, in Chapter 2 I reported a novel selection scan as well as empirical evidence that positive selection can overcome drift in heavily bottlenecked founder populations.

In Chapter 3 I described a study in which I analysed ddRADseq SNP datasets aiming to draw inferences about the demographic and evolutionary history of the native UK roe deer (*Capreolus capreolus*) population. This population, which was cut-off from mainland Europe due to rising sea levels at the start of the Holocene (i.e. 6-7 kya), was represented by roe deer samples collected from Ayrshire (Scotland). The European mainland roe deer population was represented by roe deer samples collected from Wurttemberg (Germany) and Aurignac (France). Included in the study were also samples from the introduced roe deer population in East Anglia, which was founded at the end of the 19th century, when 12 individuals were translocated from Wurttemberg to East Anglia.

Genetic distance and genetic diversity estimates indicated that, despite the Ayrshire population being isolated for ~6,000 year and the East Anglia population for less than 150 years, the East Anglia population is genetically more diverged from the mainland population and contains less segregating sites than the Ayrshire (i.e. native UK) population. Stairwayplot analyses indicated that the effective population size of the native UK roe deer population has numbered a few thousand individuals throughout the Holocene. These findings indicate moderate levels of genetic drift within the native UK roe deer population, leading to limited loss of standing genetic variation.

Whereas the selection scans FSTHet and OutFLANK did not report outliers, two SNPs were identified by both GWDS and Pcadapt as outliers potentially under positive selection in the native UK population. Alignment to the *C. pygargus* reference genome indicated that these two outlier SNPs were adjacent and 200 kB apart, and segregating in all populations. I concluded that neither genetic drift nor

diversifying selection has been of sufficient magnitude to cause fixed differences between the native UK and mainland roe deer populations, despite ~1,500 generations of isolation. I also presented a Bayesian method for population assignment.

In Chapter 4 I described a study in which I analysed whole genome sequencing data to draw inferences about the demographic and evolutionary history of the extant roe deer sister species: the European roe deer (*C. capreolus*) and the Siberian roe deer (*C. pygargus*). To infer the demographic history of these species, I used the PSMC software as well as a custom-built tool which estimates the time to the most recent common ancestor (TMRCA) based on a naive random walk Markov chain model. For selection analyses, I extracted and aligned the exomes of *C. pygargus* and *C. capreolus*, as well of those of 12 other deer species, and subsequently executed both codeml branch site tests and a custom-built tool which aims to detect genes with accelerated dN/dS rates within foreground branches.

The demographic analyses indicated a split time of of maximum 1.6 Mya – more recent than published estimates (2-4 Mya) previously inferred from mitochrondrial-DNA comparisons – and a strong difference in effective population size (Ne) throughout the separate lifespan of the sister species. The selection analyses indicated that the species with the lower historical Ne estimates, *C. capreolus*, contains higher proportions of lineage specific amino acid substitutions. Codeml branchsite tests marked 4 and 34 out of >20K genes as outlier genes in *C. pygargus* and *C. capreolus* respectively, of which the majority contained clusters of adjacent mutations in the foreground lineage.

dN/dS analyses indicated that purifying selection left a strong signature on the exomes of *Capreolus* species and of deer species in general. When ignoring the relatively minor contribution of diversifying selection, the proportion of neutral non-synonymous mutations equals the dN/dS ratio (ω), and the proportion of deleterious non-synonymous mutations equals 1 – ω (Eyre-Walker and Keightley, 2007). I found that the mean dN/dS values for various pairwise deer species comparisons range between 0.26 and 0.32 (Fig. 4.3), suggesting that approximately 70 percent of non-synonymous mutations have been purged by purifying selection.

On the power of selection scans

In summary, the selection analyses in Chapter 2 resulted in 3 out of ~80K SNPs (~0.004%) being marked as outliers, possibly being under diversifying selection. The selection analyses in Chapter 3 resulted in 2 out of ~50K SNPs (~0.004%) being marked as outliers, possibly being under diversifying selection. In Chapter 4, I found that respectively 4 and 34 out of >20K genes (~0.02% and ~0.17%) were marked by codeml branch site tests as positively selection genes (PSGs) in *C. pygargus* and *C. capreolus* respectively. After exclusion of genes which contained multinucleotide mutation clusters, which have been shown to cause false inference of positive selection(Venkat et al., 2018), the number of PSGs went down to 2 (0.01%) and ~6 genes (0.03%) respectively.

The observed proportions of codeml outlier genes (~0.02% and 0.17%) falls within the range reported by previous studies (see discussion Chapter 4, and references within). In contrast, the observed proportions of outlier SNPs fall slightly below the range reported in other genome wide selection analyses studies, with proportions of outlier SNPs ranging from 0.02% to 7.6%, with a median around 1.0% (see Appendix 1A, and references within). The relatively small size of the outlier SNP subsets presented in this thesis might reflect a conversative approach. I required SNPs to be marked by multiple selection scans in order to be considered true outliers.

The obtained proportions of SNPs and genes which have possibly experienced positive selection, seems consistent with the neutral theory, which holds that the majority of differences between populations and species are driven by neutral substitutions (Kimura, 1991). But how reliable are the obtained estimates? Does a low number of outliers indicate the absence of adaptive loci or instead a high false negative rate? (Weigand and Leese, 2018).

For my SNP datasets (Chapters 2 and 3), I answered this question by supporting the empirical data analysis with simulations which assess the power and specificity of selection scans under the given demographic settings and study design settings. Several simulation studies compare the performance of SNP based selection scans (e.g. Lotterhos and Whitlock, 2014; Luu et al., 2017) but these studies assess the performance of selection scans under a limited number of scenarios, and the results are difficult to extrapolate to specific study systems.

The simulations in Chapter 2 and Chapter 3 generated estimates of the power of selection scans under various combinations of effective population size (Ne) and selection coefficient magnitude (s) for populations with a TMRCA of 20 generations and a sample size of 30 individuals per population. These simulations suggested that due to the workings of genetic drift – which causes elevated levels of background neutral loci which make it harder for adaptive loci to stand out – the majority of positively selected loci within the South Georgia populations (Chapter 2) can not be detected by the selection scans (Fig 2.5, 2.6C), implying potentially higher numbers of positively selected regions than detected by our selection scans. The simulations also indicated that SNPs marked as outliers by multiple selection scans are likely true loci under selection. For a demographic scenario which resembles the demographic history of the native UK roe deer population (Chapter 3), the simulations indicated that almost all loci under positive selection ($s \geq 0.01$) are detected by GWDS (Fig 3.9). If the assumptions of the simulation model hold true, it is unlikely that loci under positive selection were overlooked.

Concerns have been raised about the performance of the codeml branch site test (Nozawa et al., 2009), but subsequent simulation studies have confirmed that the branch site test is generally a robust test with low false positive and false negative rates (Diekmann and Pereira-Leal, 2016; Gharib and Robinson-Rechavi, 2013; Yang and dos Reis, 2011) as long as the proportion of missing data is low (Yang and dos Reis, 2011), the number of species within the dataset sufficiently high (Delsuc and Tilak, 2015), and the proportion of selected codons within a gene equal or above 0.1 (Yang and Reis, 2011). However, violation of this latter condition might occur frequently, and lower proportions of codons under selection are associated with high (>0.8) false negative rates.

I did not test the performance of codeml branch site tests in the context of *C. capreolus* and *C. pygargus* demographies, and neither evaluated how inclusion or exclusion from other cervid species affected the outcome. Simulation studies indicate that the composition of the species tree affects the outcome of the codeml branch site test (Diekmann and Pereira-Leal, 2016). Hence, it is uncertain how well the codeml branch site test performs for the gene alignments presented in this study.

Recently it has been found that multi nucleotide mutations (MNM) cause branch site tests to incorrectly infer positive selection (Venkat et al., 2018) and indeed I found that the majority of genes marked as PSG's by codeml branch site tests contained clusters of adjacent mutations in the foreground branches. Exclusion of PSGs with MSMs reduced the number of outlier genes from 4 and 34 to respectively ~2 genes (0.01%) and ~6 genes (0.03%).

On the potential number of episodic positive selection events

The theory of the cost of natural selection holds that selection can act on a limited number of adaptive loci at a time, due to mortality costs associated with substitution events (Haldane, 1957). Haldane argued that fixation time is proportionally related to the number of adaptive loci: if it takes *t* generations to fixate an adaptive allele at one locus, it should take *Lt* generations to fixate adaptive alleles at *L* loci (Hickey and Golding, 2019). Although simulations do not support these theoretical constraints (Nunney, 2003), it still seems intuitive and reasonable to assume that if multiple adaptive alleles are present within a population, chances are they occur in different individuals and therefore will compete against each other, slowing down the adaptation process by prolonging fixation times (Weissman and Barton, 2012). A recent simulation study however indicates that the frequency of adaptive alleles can respond simultaneously at many loci to independent selection at rates similar to the predicted rate for single locus selection (i.e. within several hunderd generations given a selection coefficient of 0.02, Hickey and Golding, 2019). This finding suggests that natural selection can drive many adaptive alleles to fixation within ecological time scales.

Given the age of the South Georgia reindeer populations (~100 years or ~20 generations), limited time has been available for natural selection to drive alleles to fixation. If assuming that the average effective population size of the South Georgia founder populations equalled 25 individuals, then the fixation time of a neutral alleles averages 100 generations (4*Ne, Kimura and Crow, 1964). Soft sweeps require less generations to complete, because selection speeds up the fixation process and also because the original founders might carry multiple copies of the adaptive allele, but likely not less than 20 generations. It can therefore be argued that the observed differences in minor allele frequencies of the outlier SNPs between

the South Georgia reindeer populations and their source population are consistent with expectations for a locus under strong diversifying selection regime, as the age of the founder populations have not been sufficient to drive alleles to fixation.

Evaluation of the outcome of the selection analyses on UK and mainland roe deer populations (Chapter 3) leads to a different conclusion. Assuming an average generation time of 5 years (Nilsen et al., 2009) and assuming that the UK roe deer got cut-off from the mainland population around 6-7 kya (Coles, 1998; Sturt et al., 2013), the UK roe deer population has been isolated for at most 1500 generations. Stairway plot analyses indicate that the effective population size of the native UK roe deer population equalled approximately 5000 individuals throughout the Holocene (Fig 3.7). Since roe deer are provincial (Baker and Rus Hoelzel, 2012) and since all native UK roe deer samples analysed in Chapter 3 derived from Ayrshire, fixation of alleles within this local population might require a less extensive sweep. In either case, there has been ample time for positive selection to drive adaptive alleles to fixation, at many loci. Despite this potential, I only found one genomic locus to be possibly under diversifying selection, represented by SNPs which are segregating in all three study populations (i.e. no fixed differences). This finding suggests the near absence of genomic regions (and hence phenotypic traits) which differential fitness effects between the UK and mainland roe deer populations in the past 6-7 ky, and furthermore that the only potential exception – the genomic region which harbours the two outlier SNPs – has been under very weak selection at most.

Similarly, the observed number of positively selected genes (PSGs) in *C. pygargus* and *C. capreolus* (Chapter 4) lags far behind the potential number of PSGs. Although the TMRCA of these species is more than 1 Mya, less than 10 genes contain codon substitutions driven by positive selection in either of the two species. These PSGs contained each just a few non-synonymous substitutions driven by positive selection, out of a total of 170,596 exomic single nucleotide differences between the two species.

On the difference between observed and potential number of positive selection

What explains the apparant difference between the observed number of SNPs/genes under diversifying selection and the potential number of SNPs/genes based on theoretical expectations? Potential explanations are:

i. Low power of selection scans
ii. Genomic regions/features targetted by diversifying selection are not represented in the datasets
iii. Selection is not pervasive but episodic
iv. Absence of diversifying selection

The first explanation – low power of selection scans – is not supported by simulations studies (but see explanation iii). Simulations in Chapter 2 and Chapter 3 of this thesis indicated high power of GWDS in the context of the demographic scenario of the native UK roe population (Fig. 3.9). Similarly, simulation studies suggest that the codeml branch site test generally has low false negative rates (Diekmann and Pereira-Leal, 2016). However, violation of assumptions of the simulation models might affect the outcome. The same is true for data artifacts, such as genotyping and alignment errors.

The second potential explanation is that the datasets analysed in this thesis do not include the genomic loci and/or features under selection. The extent to which genome wide selection scans screen the entire genome depends on the density of the SNP catalogue as well as to the level of linkage disequilibrium within the study population. The higher the number of SNPs, the higher the probability that high proportions of linkage blocks are represented by one or more SNPs. The absence of stacked outlier SNPs in Manhattan plots (Fig A3.9) is suggestive of sparse sampling of genome wide genetic variation.

The second potential explanation – which holds that the datasets analysed in this thesis do not include the genomic loci and/or features under selection – might also account for the observed low number of PSGs in *C. capreolus* and *C. capreolus* (Chapter 4), and might suggest that the divergence between these species is not driven by changes within genes, but instead by other changes within the genome (Hughes, 2007), such as gene copy number variations (Rinker et al., 2019), mutations in regulatory sequences (Brawand et al., 2014; King and Wilson, 1975;

Sackton et al., 2019), *de novo* gene evolution (Baalsrud et al., 2018) and gene silencing through genomic translocations (Hof et al., 2016).

A third potential explanation for the discrepancy between the observed and potential number of adaptive changes is that selection is episodic rather than pervasive, and that this temporal variation in the magnitude and direction of the selection coefficient might complicate detection of positive selection. The fluctuating nature of selection coefficients is demonstrated by two of the most well-studied cases of contemporary evolution: industrial melanism of peppered moths (Cook and Saccheri, 2012) and beak morphology changes in Darwin finches (Weiner, 1994). Fst-based selection scans implicitly assume that positive selection acts for a sufficient period of time in the same direction in order for alleles to stand out from the background of neutral variation and to eventually cause permanent fixation in the population on which positive selection is acting. However, environmental conditions fluctuate continuously, meaning that the assumption of fixed and directional selection is routinely violated. Extreme examples are provided by the study cases of the peppered methods and the Darwin finches, in which the effects of positive selection are erased due following fluctuations in environmental changes. These examples illustrate that limited time windows may exists for selection scans to detect a positive selection event. The power and specificity estimates presented in Chapter 2 are generating using simulations assuming a fixed selection coefficient, and therefore provide no insight into the detectability of adaptive SNPs under a regime with fluctuating selection pressures.

Temporal variations of selection coefficients may also affect false positive and false negative rates of codeml branch-site tests. Lineage specific dN/dS ratios and codeml branch-site tests can generate evidence for episodic selection events, if episodic selection is defined as directional selection experienced within a specific lineage. However, reversal of fixation of temporally adaptive non-synonymous mutations can mask fingerprints of selection. This is especially problematic if reversal occurs in a subset of foreground branches, as codeml branch-site test do not allow for substitution rate variation within foreground branches.

The fourth potential explanation for the discrepancy between the observed and potential number of adaptive changes is a scarcity of positive selection events, as predicted by the neutral theory (Kimura, 1991) and the nearly neutral theory

(Ohta, 1995). The findings presented in Chapter 4 could be argued to fit particularly well with the neutral theory, which holds that many mutations are slightly delerious. In Chapter 4 it was shown that codeml branch site tests marked considerably more genes as PSGs in the species with relatively low Ne (i.e. *C. capreolus*) compared to the species with higher Ne (*C. pygargus*). Slightly deleterious mutations are less effectively purged in small populations, and the higher number of PSGs in *C. capreolus* compared to *C. pygargus* might reflect relaxation of purifying selection (Hughes, 2007). More whole exome comparisons between closely related sister species are needed to evaluate the plausibility of the relaxed purifying selection hypothesis. In addition, studies into the deleteriousness of the observed mutation clusters in the *C. capreolus* exome could clarify whether these clusters are deleterious, neutral or adaptive by nature (see for example Feng et al., 2019).

Conclusion

In this thesis I executed selection analyses on genomic datasets of reindeer and roe deer populations. For each of the three study systems I found evidence for positive selection, including in the heavily bottlenecked South Georgia reindeer founder populations. This finding provides empirical evidence that founder populations can adapt to novel environments even in the face of pronounced genetic drift. Due to the uncertaincy of the performance of selection scans for each specific dataset and due to the reduced representation of genome wide variation by SNP and exome datasets, it is unknown how faithfully the number of outlier SNPs and outlier genes outputted by selection scans reflect the number of episodic positive selection. Caution should therefore be exercised when comparing the outcomes of selection scans to predictions of the (nearly) neutral theory. The finding that codeml branch site tests marked considerably more genes as positively selected in a species with relatively low Ne (i.e. *C. capreolus*) compared to the species with higher Ne (*C. pygargus*), is suggestive of relaxation of purifying selection.

APPENDICES CHAPTER 1

Supplementary information. ***Reported numbers of outlier SNPs in a random subset of published genome wide selection analyses studies***

Mammals and birds. PCadapt marked 59 out of 22,935 SNPs (0.25%) as outliers putatively under diversifying selection between populations of eastern coyote occuring in historical pre-1900 range and populations occurring in newly colonized habitat (Heppenheimer et al., 2018). Bayescan marked 178 SNPs out 67,000 SNPs (0.26%) as being under putative diversifying selection in grey wolf populations spread throughout Eurasia (Stronen et al., 2015). Bayescan also marked up to 140 out of 5820 SNPs (2.4%) as divergent between samples of living and diseased bottlenose dolphin (Cammen et al., 2015). For a pairwise population comparison between house finch populations sampled before and after an epizootic outbreak, Bayescan marked 4 out of 18,000 SNPs (0.02%) as outliers (Shultz et al., 2016).

Marine invertebrates. Many genome wide selection scan studies focus on marine datasets. Arlequin and Bayescan marked 112 out of 7163 SNPs (1.6%) as outliers in marine bivalves off the coast of Northern America, and these SNPs exhibited enhanced isolation by distance effect outliers (Van Wyngaarden et al., 2016). The same trend of increased isolation by distance effects were observed for 129 out of 41,159 SNPs (0.31%)which were marked as outliers by at least two selection scans (among them OutFLANK, Bayescan and PCadapt, (Silliman, 2019). 34 out of 55,409 SNPs (0.06%) were marked by both Bayescan and PCadapt in closely related populations among the west coast of South Africa (Nielsen et al., 2018). 44 out of 5,484 SNPs (0.8%) were marked by both Bayescan and Arlequin in coral reefs populations subjected to an environmental (temperature) gradients along the west coast of Australia (Thomas et al., 2017).

Fish. Arlequin (Excoffier and Lischer, 2010) marked at most 139 out of 6,167 SNPs (2.3%) as putatively being under diversifying selection in Atlantic salmon populations (Bourret et al., 2013) and 59 out of 3737 SNPs (1.6%) as putatively under divergent selection in the reef fish occuring around Marquesas islands, which

split from the widespread pacific reef fish form around 0.5 Mya (Gaither et al., 2015). In another study on reef fish populations occurring in the Indian Ocean, Arlequin and Bayescan marked 26 out of 1174 SNPs (2.2%) as outliers (Salas et al.)

Arlequin and Bayescan marked 17 out of 381 SNPs (4.5%) as putatively under diversifying selection between Atlantic and Mediterrean hake (Milano et al., 2014) and 47 out of 13,674 SNPs (0.34%) putatively under diversifying selection between Red Sea and Mediterranean cornet fish, the latter having colonized the Mediterrean Sea in the year 2000 (Bernardi et al., 2016). Arlequin marked 150 out of 4439 SNPs (3.4%) as outliers in data on lamprey populations in rivers in North America (Hess et al., 2013) and PCadapt marked 88 out of 1153 SNPs (7.6%) as outliers in Mediterrean striped red mullet (Dalongeville et al., 2018). A study comparing redband trout populations occurring in desert and montane streams, resulted in 821 (0.16%), 973 (0.19%) and 865 (0.16%) out of 526,301 SNPs being marked as outliers under putative diversifying selection by respectively Bayescan, OutFLANK, and PCadapt, of which 435 SNPs (0.08%) were identified by at least two scans (Chen et al., 2018).

Plants. Bayescan also marked up to 38 loci out of 15,000 SNPs (0.25%) as outliers for comparisons between populations representing various plant ecotypes (i.e. populations occurring on beaches, in estuaries and springs) in Scandanivia (Brandrud et al., 2017).

APPENDICES CHAPTER 2

Table A2.1. Sequencing output. *STACKS demultiplexing output.*

Sequencing Date	ID	Total Reads	No RadTag	Low Quality	Retained	Read pairs	U fod	U rev	Barcode	Pool
Dec-15	6	3953892	21639	10271	3921982	1952198	8121	9465	GCATG	Pool1
Dec-15	10	898478	8172	2082	888224	442833	1283	1275	AATCG	Pool1
Dec-15	13	3080760	12551	8396	3059813	1526009	6252	1543	ACAGA	Pool1
Dec-15	18	6480662	7916	16414	6456332	3223980	5881	2491	AAGTGA	Pool1
Dec-15	19	2819058	3769	6418	2808871	1402746	2374	1005	ATTACA	Pool1
Dec-15	24	758834	2766	1804	754264	376389	1155	331	CAGGCG	Pool1
Dec-15	25	9746726	10276	23994	9712456	4849721	9802	3212	AGAATGA	Pool1
Dec-15	26	13263556	12116	41212	13210228	6596466	12849	4447	AGTTAAT	Pool1
Dec-15	27	11001960	42074	44049	10915837	5439714	25115	11294	GCATG	Pool2
Dec-15	28	3233724	16879	13333	3203512	1598019	4818	2656	AATCG	Pool2
Dec-15	29	7048322	24361	29762	6994199	3488312	14412	3163	ACAGA	Pool2
Dec-15	30	10137206	23899	44833	10068474	5021636	20126	5076	AAGTGA	Pool2
Dec-15	32	7119006	17401	27918	7073687	3527733	15475	2746	ATTACA	Pool2
Dec-15	33	3642716	8784	15836	3618096	1805048	6230	1770	CAGGCG	Pool2
Dec-15	34	8115326	17280	23779	8074267	4027557	16275	2878	AGAATGA	Pool2
Dec-15	36	11146988	26766	37437	11082785	5525879	26332	4695	AGTTAAT	Pool2
Dec-15	37	9161318	38562	24875	9097881	4531677	22615	11912	GCATG	Pool3
Dec-15	41	2747378	14390	7946	2725042	1358601	5425	2415	AATCG	Pool3
Dec-15	42	5535034	12851	73426	5448757	2716518	11320	4401	ATTACA	Pool3
Dec-15	43	2409390	5511	31003	2372876	1183611	3068	2586	CAGGCG	Pool3
Dec-15	46	9481348	22091	19772	9439485	4707538	21655	2754	AGAATGA	Pool3
Dec-15	48	9001682	32809	21393	8947480	4455849	32842	2940	AGTTAAT	Pool3
Dec-15	49	7198404	30050	29337	7139017	3556645	14177	11550	GCATG	Pool4
Dec-15	50	1576452	10663	6636	1559153	777475	2551	1652	AATCG	Pool4
Dec-15	51	7183866	20763	29968	7133135	3558224	13787	2900	ACAGA	Pool4
Dec-15	52	6941282	12986	29788	6898508	3442228	10841	3211	AAGTGA	Pool4
Dec-15	53	4376460	9375	17468	4349617	2169946	8109	1616	ATTACA	Pool4
Dec-15	54	2521742	5825	10108	2505809	1250284	4170	1071	CAGGCG	Pool4
Dec-15	55	8743296	16175	27377	8699744	4340252	15935	3305	AGAATGA	Pool4
Dec-15	56	6536100	10780	21825	6503495	3245047	10704	2697	AGTTAAT	Pool4
Dec-15	57	5363876	18364	14649	5330863	2657800	9928	5335	GCATG	Pool5
Dec-15	61	2236470	9119	5896	2221455	1108568	2232	2087	AATCG	Pool5
Dec-15	62	5635602	15088	16334	5604180	2796177	10184	1642	ACAGA	Pool5
Dec-15	65	7065040	14521	21739	7028780	3506673	12830	2604	AAGTGA	Pool5
Dec-15	66	5268192	12969	15126	5240097	2613206	12119	1566	ATTACA	Pool5
Dec-15	69	8932100	25721	25344	8881035	4429566	19244	2659	ACAGA	Pool3
Dec-15	72	6869500	78453	30067	6760980	3340909	73350	5812	AAGTGA	Pool3
Dec-15	74	2000950	3920	5478	1991552	994161	2574	656	CAGGCG	Pool5
Dec-15	77	4458660	12138	8670	4437852	2212522	11522	1286	AGAATGA	Pool5
Dec-15	79	5720182	9823	12194	5698165	2843361	9796	1647	AGTTAAT	Pool5

Dec-15	N15	4341730	3140	12308	4326282	2160768	3186	1560	CCACTGG	Pool1
Dec-15	N16	11027632	26852	38668	10962112	5465133	25972	5874	CCACTGG	Pool2
Dec-15	N26	4655634	12238	10900	4632496	2309372	12119	1633	CCACTGG	Pool3
Dec-15	N30	2707412	9956	10022	2687434	1338159	9436	1680	CCACTGG	Pool4
Dec-15	N34	5592944	11194	11624	5570126	2778550	11263	1763	CCACTGG	Pool5
Dec-15	N35	2552716	4314	6536	2541866	1268353	2688	2472	AGTCAAGA	Pool1
Dec-15	N36	3063728	10309	9803	3043616	1516053	8836	2674	AGTCAAGA	Pool2
Dec-15	N37	4822116	16382	11120	4794614	2388237	14981	3159	AGTCAAGA	Pool3
Dec-15	N38	2288548	7709	8952	2271887	1131537	6691	2122	AGTCAAGA	Pool4
Dec-15	N39	5005648	10857	10837	4983954	2485718	10228	2290	AGTCAAGA	Pool5
Dec-15	N40	2141908	3690	5297	2132921	1064364	3072	1121	AGTGTTAA	Pool1
Dec-15	N41	5655788	15176	17566	5623046	2802919	14162	3046	AGTGTTAA	Pool2
Dec-15	N42	6583760	16454	14186	6553120	3267241	15865	2773	AGTGTTAA	Pool3
Dec-15	N43	4061154	10538	13843	4036773	2012171	10086	2345	AGTGTTAA	Pool4
Dec-15	N44	1147684	3685	2387	1141612	568893	3230	596	AGTGTTAA	Pool5
Dec-15	N45	3845336	5682	9353	3830301	1911662	5756	1221	CACGACCA	Pool1
Dec-15	N46	4620888	12668	13237	4594983	2290253	12584	1893	CACGACCA	Pool2
Dec-15	N48	1331872	4301	3337	1324234	659834	4030	536	CACGACCA	Pool3
Dec-15	N49	3200932	9538	11527	3179867	1584464	9366	1573	CACGACCA	Pool4
Dec-15	N50	2252732	6976	4614	2241142	1116784	6837	737	CACGACCA	Pool5
Jun-16	7	5665606	64094	60515	5540997	2751705	19824	17763	GCATG	Pool1
Jun-16	8	1110248	39822	10620	1059806	525417	4225	4747	AATCG	Pool1
Jun-16	14	8309212	76807	115745	8116660	4032374	42389	9523	ACAGA	Pool1
Jun-16	15	10168782	54867	126152	9987763	4967964	40939	10896	AAGTGA	Pool1
Jun-16	16	8424024	65368	127965	8230691	4082654	53071	12312	ATTACA	Pool1
Jun-16	20	808404	11515	7541	789348	393024	2146	1154	CAGGCG	Pool1
Jun-16	21	9542184	27626	106757	9407801	4689072	20657	9000	AGAATGA	Pool1
Jun-16	22	9343570	22366	121077	9200127	4586363	18622	8779	AGTTAAT	Pool1
Jun-16	23	5465414	13069	64248	5388097	2686066	10554	5411	CCACTGG	Pool1
Jun-16	35	6959646	24699	99032	6835915	3402345	19380	11845	AGTCAAGA	Pool1
Jun-16	38	10567088	82615	167703	10316770	5112011	71456	21292	AGTGTTAA	Pool1
Jun-16	39	10477990	46825	119327	10311838	5128782	42233	12041	CACGACCA	Pool1
Jun-16	40	3472424	61027	69987	3341410	1655747	18647	11269	GCATG	Pool2
Jun-16	67	916990	49192	14701	853097	421572	3845	6108	AATCG	Pool2
Jun-16	68	6058844	63591	103262	5891991	2929340	24653	8658	ACAGA	Pool2
Jun-16	70	10706444	53490	175673	10477281	5212854	37216	14357	AAGTGA	Pool2
Jun-16	71	7759316	45168	139865	7574283	3764860	33004	11559	ATTACA	Pool2
Jun-16	91	335224	11285	4723	319216	158605	879	1127	CAGGCG	Pool2
Jun-16	94	2897684	24962	134209	2738513	1356428	16745	8912	AGAATGA	Pool2
Jun-16	98	5579306	19839	92442	5467025	2722395	15000	7235	AGTTAAT	Pool2
Jun-16	101	6556984	19259	101441	6436284	3205953	15489	8889	CCACTGG	Pool2
Jun-16	102	8974316	23939	143990	8806387	4385731	19070	15855	AGTCAAGA	Pool2
Jun-16	105	14200126	43968	230603	13925555	6932319	38572	22345	AGTGTTAA	Pool2
Jun-16	58	7650832	19265	118791	7512776	3742623	17245	10285	CACGACCA	Pool2
Jun-16	59	5298956	48177	124175	5126604	2548429	12097	17649	GCATG	Pool3
Jun-16	60	1043490	35260	20752	987478	489823	2655	5177	AATCG	Pool3

Jun-16	63	6869100	43450	166632	6659018	3316230	14849	11709	ACAGA	Pool3
Jun-16	64	9719836	31937	239994	9447905	4704244	21090	18327	AAGTGA	Pool3
Jun-16	75	8939390	30477	211218	8697695	4329658	22073	16306	ATTACA	Pool3
Jun-16	80	705486	9431	14704	681351	338895	1849	1712	CAGGCG	Pool3
Jun-16	82	7821014	22217	181080	7617717	3792667	16520	15863	AGAATGA	Pool3
Jun-16	81	5324900	14361	129594	5180945	2579449	10973	11074	AGTTAAT	Pool3
Jun-16	84	4505818	11530	103603	4390685	2186021	9119	9524	CCACTGG	Pool3
Jun-16	85	8032706	15741	194957	7822008	3894573	13579	19283	AGTCAAGA	Pool3
Jun-16	86	10376158	29672	259746	10086740	5017433	25258	26616	AGTGTTAA	Pool3
Jun-16	88	5304834	13697	122955	5168182	2572126	12304	11626	CACGACCA	Pool3
Jun-16	89	4992548	43635	32970	4915943	2448028	11845	8042	GCATG	Pool4
Jun-16	90	1076870	32789	7054	1037027	515448	2796	3335	AATCG	Pool4
Jun-16	N17	9255604	55772	70773	9129059	4547055	27228	7721	ACAGA	Pool4
Jun-16	N51	10778820	38834	90271	10649715	5306017	27176	10505	AAGTGA	Pool4
Jun-16	N52	10602902	35022	83733	10484147	5223997	27074	9079	ATTACA	Pool4
Jun-16	N53	711882	9925	5139	696818	346833	2127	1025	CAGGCG	Pool4
Jun-16	N55	4029460	15878	32165	3981417	1983363	10523	4168	AGAATGA	Pool4
Jun-16	N57	7458572	18224	58318	7382030	3680180	14884	6786	AGTTAAT	Pool4
Jun-16	N59	4390230	11112	35004	4344114	2165260	9188	4406	CCACTGG	Pool4
Jun-16	N64	2543196	7348	22841	2513007	1251778	5455	3996	AGTCAAGA	Pool4
Jun-16	N66	14264594	34227	111523	14118844	7036282	30115	16165	AGTGTTAA	Pool4
Jun-16	N67	5620068	14069	40965	5565034	2773369	12891	5405	CACGACCA	Pool4
Jun-16	N68	5263906	39625	139445	5084836	2529099	11991	14647	GCATG	Pool5
Jun-16	N69	1373022	30596	35130	1307296	649320	3840	4816	AATCG	Pool5
Jun-16	N71	5236678	43263	153825	5039590	2504763	19364	10700	ACAGA	Pool5
Jun-16	N74	6460676	29189	190103	6241384	3103354	19762	14914	AAGTGA	Pool5
Jun-16	N75	7810984	31105	223119	7556760	3757823	24442	16672	ATTACA	Pool5
Jun-16	N80	653046	8038	20784	624224	310283	1750	1908	CAGGCG	Pool5
Jun-16	N81	4449352	14121	122045	4313186	2146732	9400	10322	AGAATGA	Pool5
Jun-16	N83	6740164	16937	203042	6520185	3244698	14122	16667	AGTTAAT	Pool5
Jun-16	N84	2866206	6623	97831	2761752	1374362	5153	7875	CCACTGG	Pool5
Jun-16	N85	4031692	8194	141151	3882347	1931266	6744	13071	AGTCAAGA	Pool5
Jun-16	N87	10802312	23616	337899	10440797	5193544	20505	33204	AGTGTTAA	Pool5
Jun-16	N89	5088996	14385	143432	4931179	2452429	13054	13267	CACGACCA	Pool5
Sum		692705826	2754390	7704415	682247021	339812578	1787522	834343		
% total reads		100	0.4	1.1	98.5	98.1	0.3	0.1		
Average		5772549	22953	64203	5685392	2831771	14896	6953		
Stdev		3257785	17545	69668	3213602	1600722	12542	6227		

Table A2.2. SNP dataset summary statistics

	Before filtering	After filtering	After thinning
Number of individuals	120	95	95
Number of SNPs	87876	67481	27690
Percentage of SNPs with maf >= 0.05	53.66	66.09	65.56
Mean spacing between SNPs	23374.05	23326.16	60302.78
Median spacing between SNPs	229	235	38024
Mean proportion of missing data per individual	0.17	0.04	0.04
GC content	0.61	0.6	0.61
Transition vs transversion ratio	1.96	2.08	2.16

Table A2.3. Bayesass3-SNPs migration rates

	Busen	Barff	Norway
Busen	0.9796(0.0139)	0.0100(0.0099)	0.0103(0.0100)
Barff	0.0097(0.0095)	0.9806(0.0131)	0.0097(0.0094)
Norway	0.0092(0.0090)	0.0091(0.0088)	0.9817(0.0122)

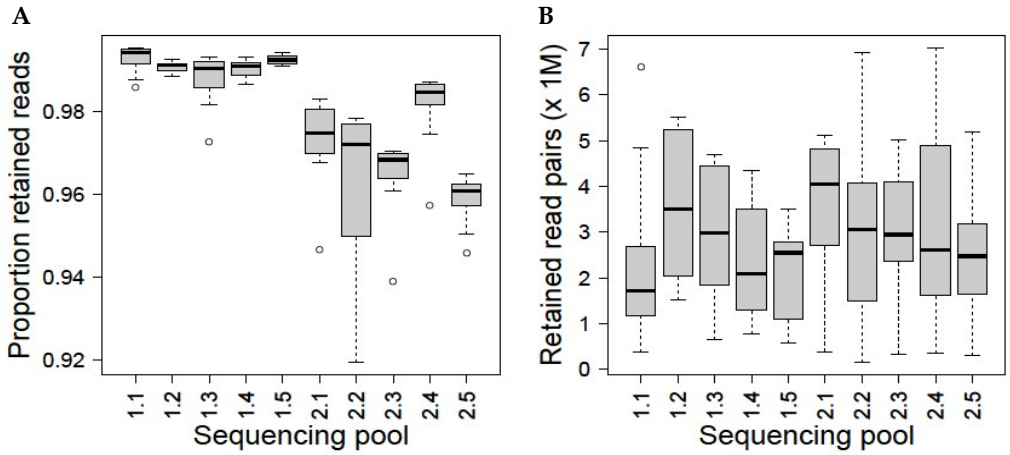


Fig. A2.1. Retained reads per sample. A. *Proportion of retained read pairs after removing low quality reads and reads with missing radtag of missing mate pair.* ***B.*** *Number of retained read pairs per sample.*

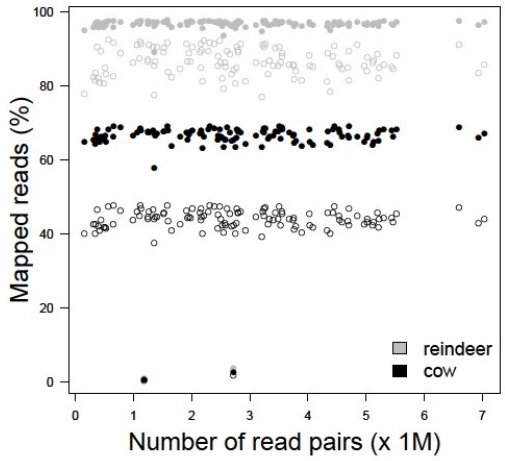


Fig. A2.2 Sample specific alignment rates*. Closed circles: all alignments. Open circles: concordant alignments. Black: alignment to cow genomes. Grey: alignment to reindeer genome.*

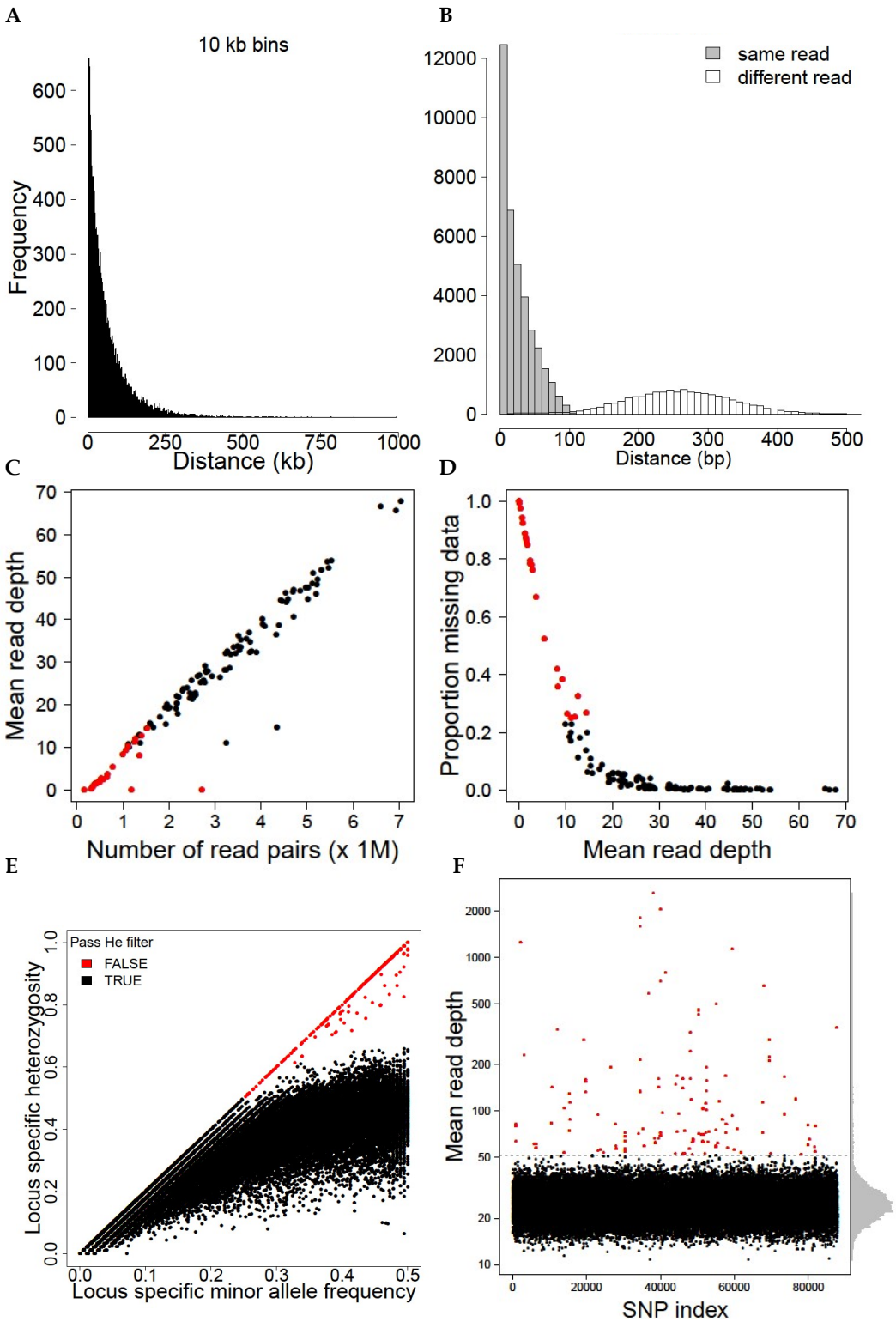


Fig. A2.3. SNP spacing and data quality control*. Spacing between SNPs (A) and quality control assessment (B-D). For B-D: black indicates retained loci/samples and red indicates filtered loci/samples.* ***A-B****. Spacing between SNPs.* ***C.*** *Sample specific read depth versus number of retained read pairs per sample .* ***D.*** *Missing data per sample versus mean read depth per sample.* ***E.*** *Locus specific heterozygosity versus locus specific minor allele frequency. Excessive heterozygosity excess is indicative of paralogous loci.* ***F.*** *Mean read depth per locus.*

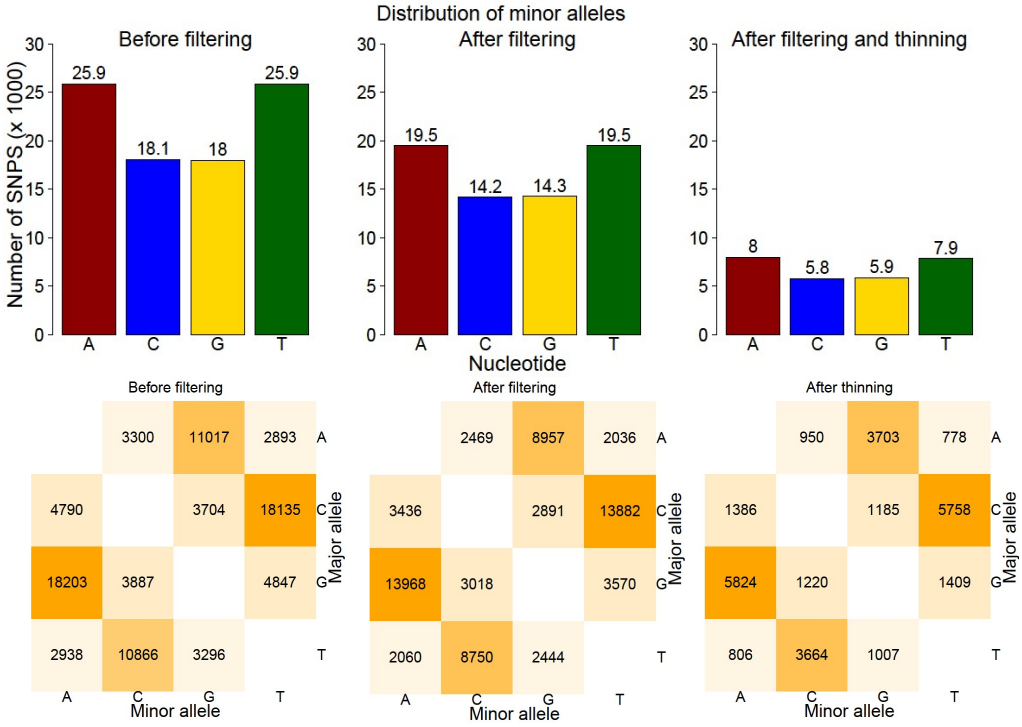


Fig.A2.4. GC content. *GC content and transition vs transition ratios for unfiltered, filtered and thinned SNP datasets.*

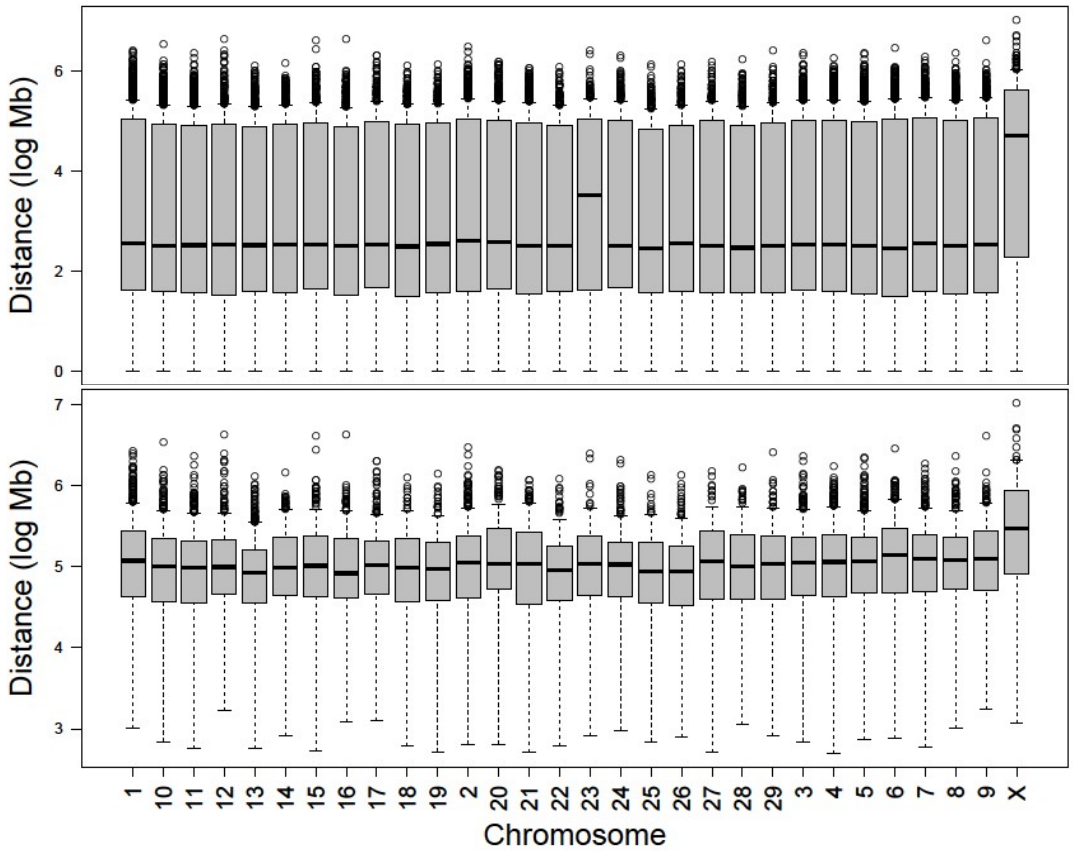


Fig.A2.5. Distribution of SNPs over chromosomes. *SNP spacing per chromosome for filtered (above) and thinned (below) SNP datasets. Estimates based on alignment against cow genome.*

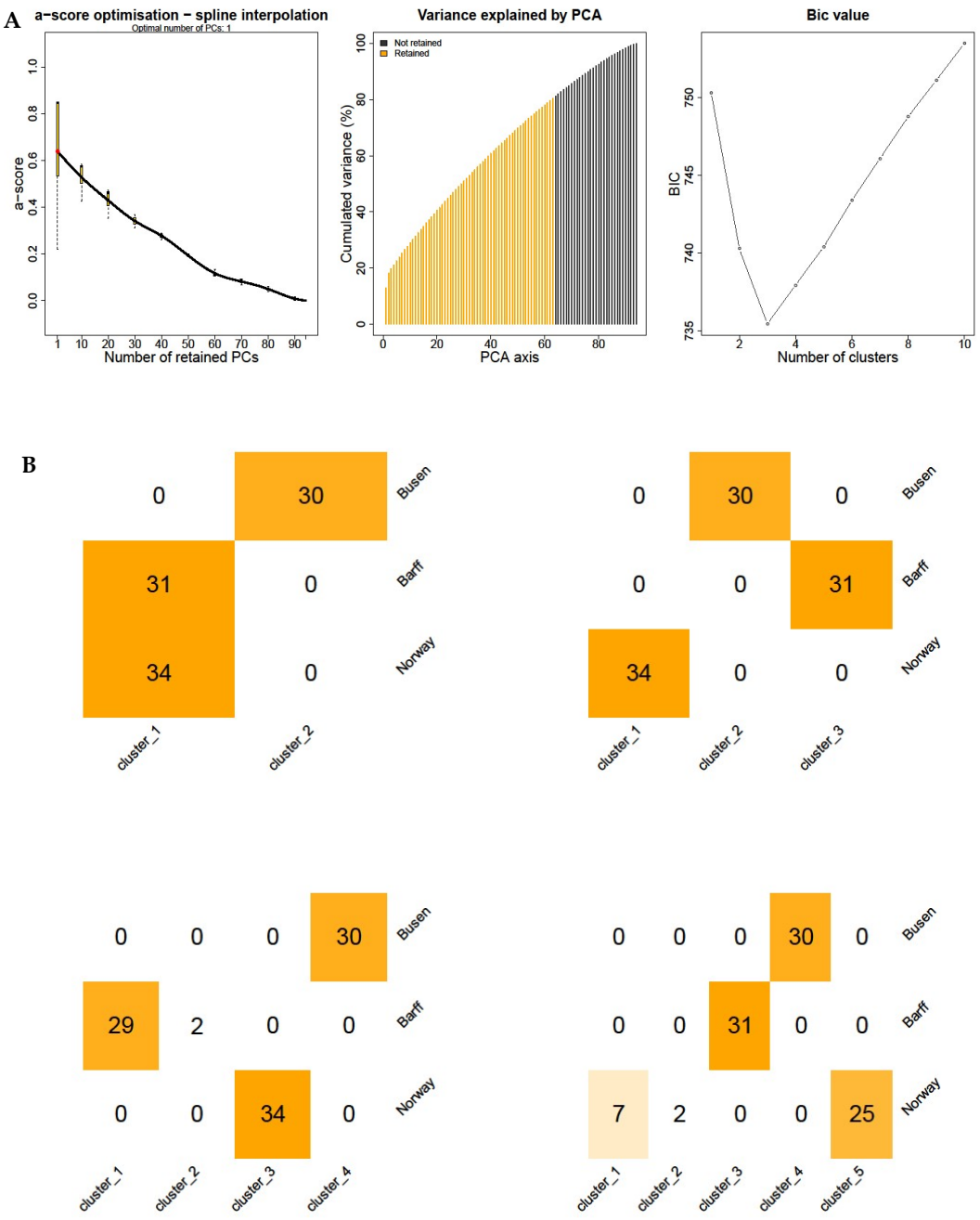


Fig.A2.6. DAPC analysis. (A) *DAPC summary statistics: a-score, number of retained PC's, and bic value. Because the 'a-score optimisation – spline interpopulation'-method returned an optimum number of 1 retained PCs, I opted for another approach and selected a number of PCs which explained 80 percent of cumulated variance.* ***(B).*** *Expected population clustering (Busen, Barff, Norway) vs DAPC inferred clusters for K = 2-5.*

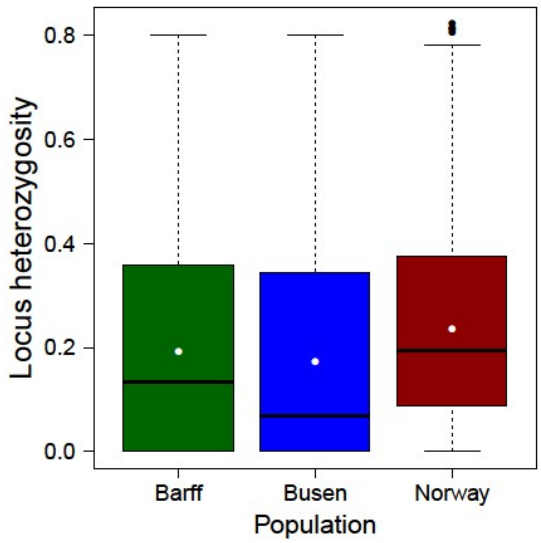


Fig.A2.7. Boxplots of locus specific heterozygosity per population. *White dots indicate means. Based on filtered and thinned dataset.*

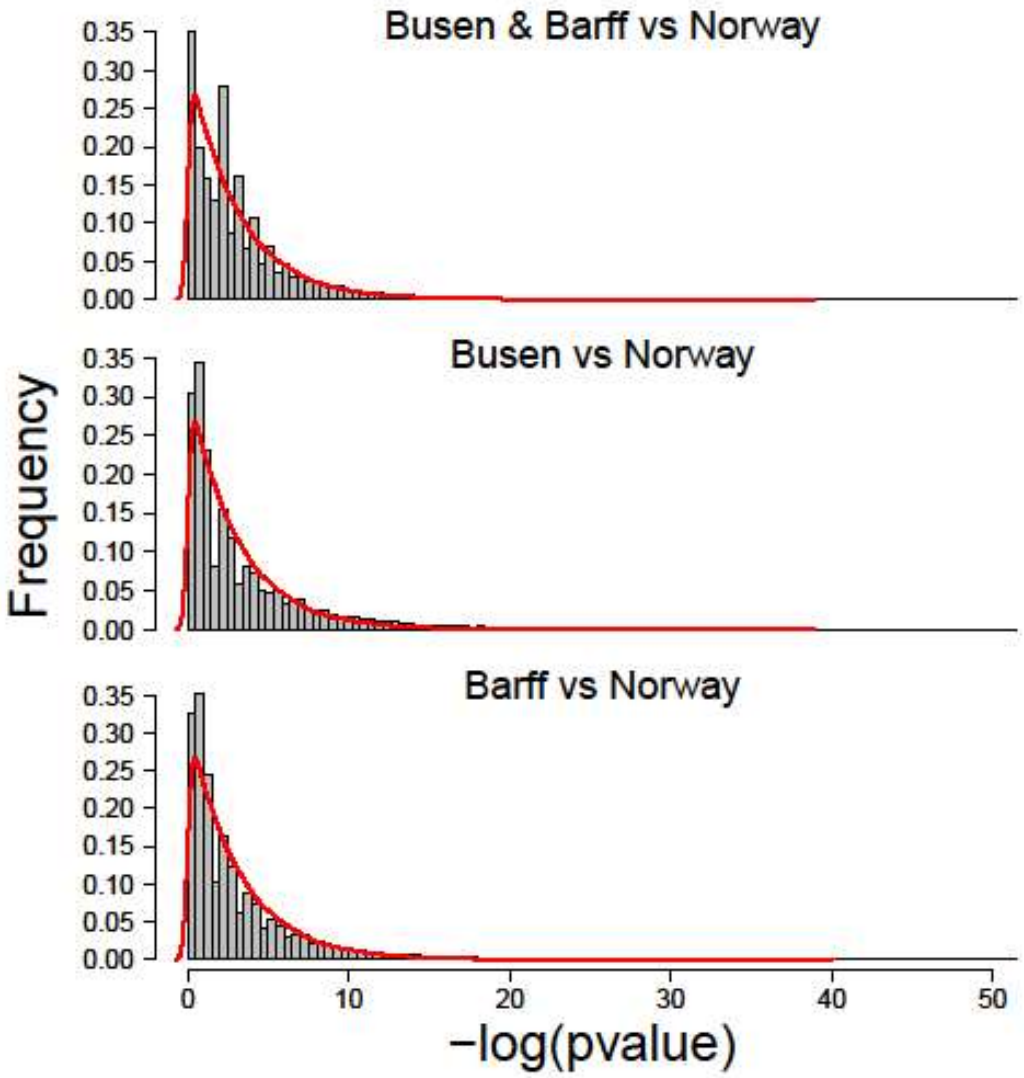


Fig.A2.8A. Distribution of negative natural log of GWDS fisher exact test scores. *Grey bars: observed distribution. Red lines: exponentional distributions fitted to the data by GWDS.*

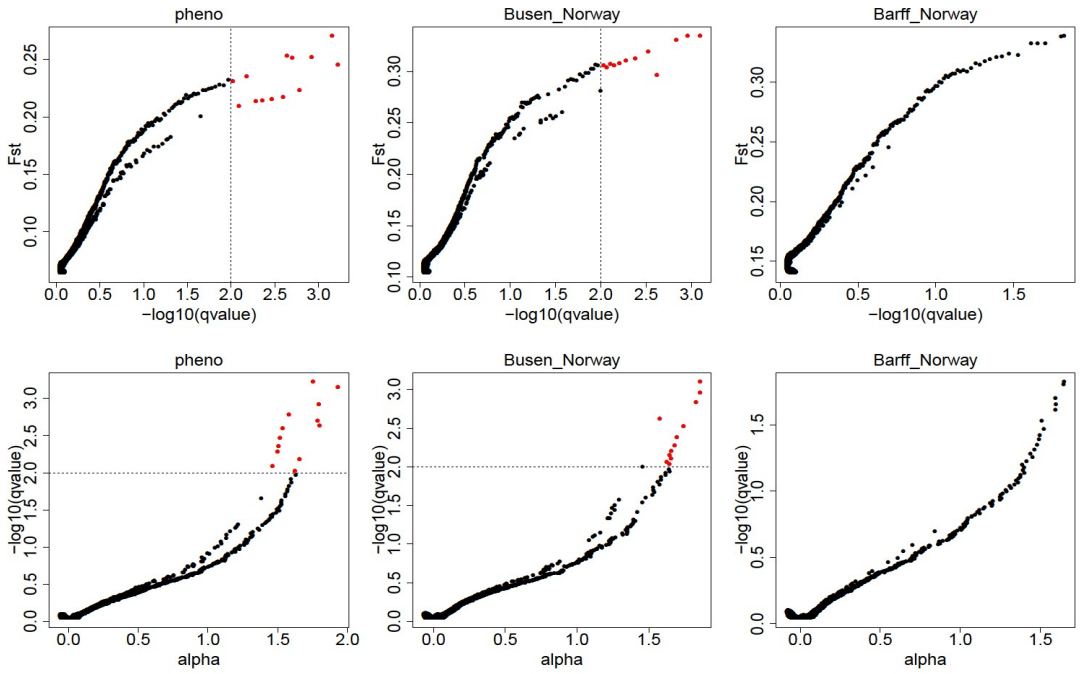


Fig.A2.8B. Bayescan test results. *Red scores are loci scored as outliers by Bayescan with a false discovery rate of 0.1. All candidate outlier loci have a positive alpha value, indicating that none of the putatitive outliers are under purifying/balancing selection, but instead under positive selection.*

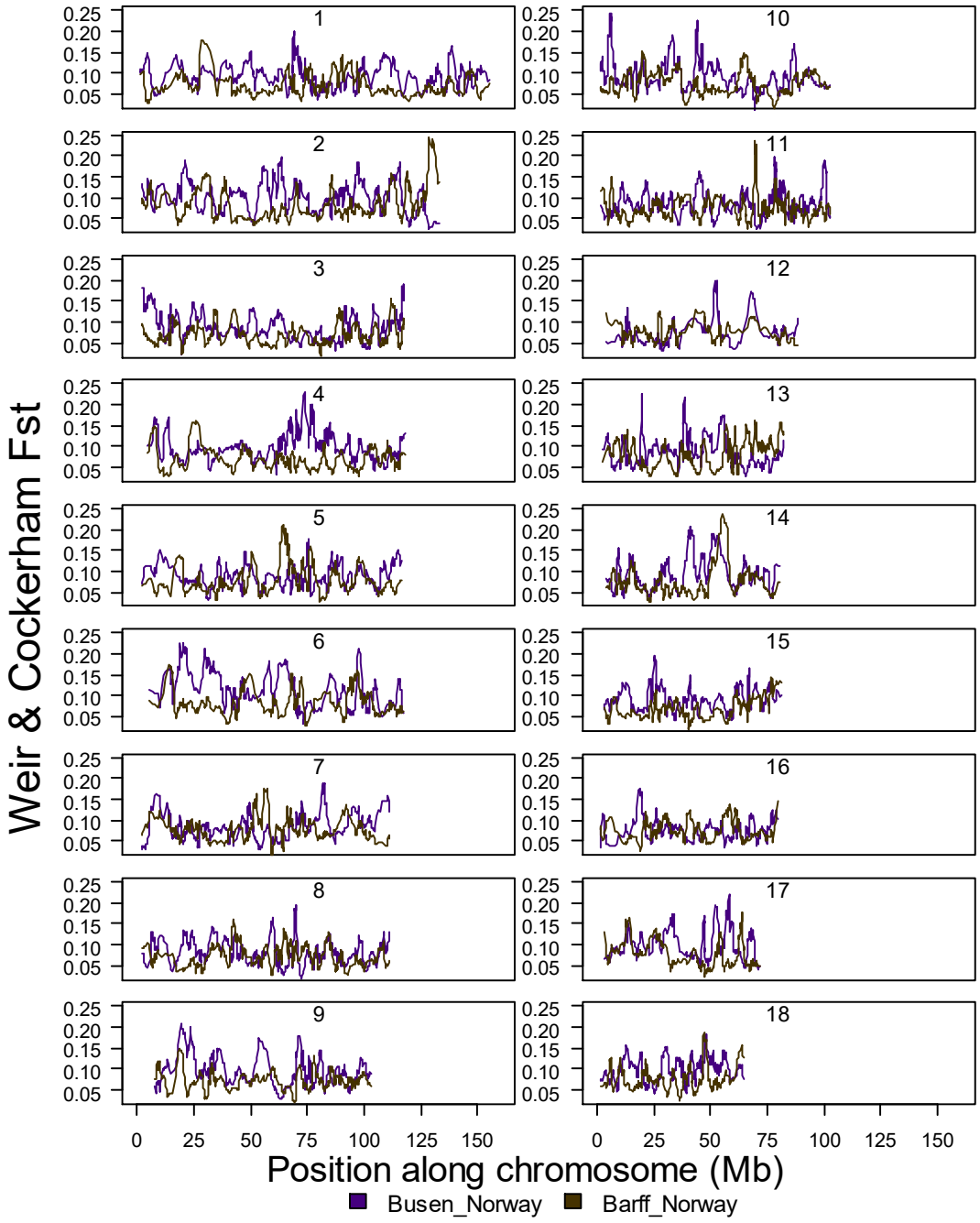


Fig.A2.9. Sliding window Fst. *Sliding window Weir & Cockerham Fst plots for pairwise population comparisons between both founder populations and their source population. Window size is 20 SNPs, equalling approximately 0.5 Mb.*

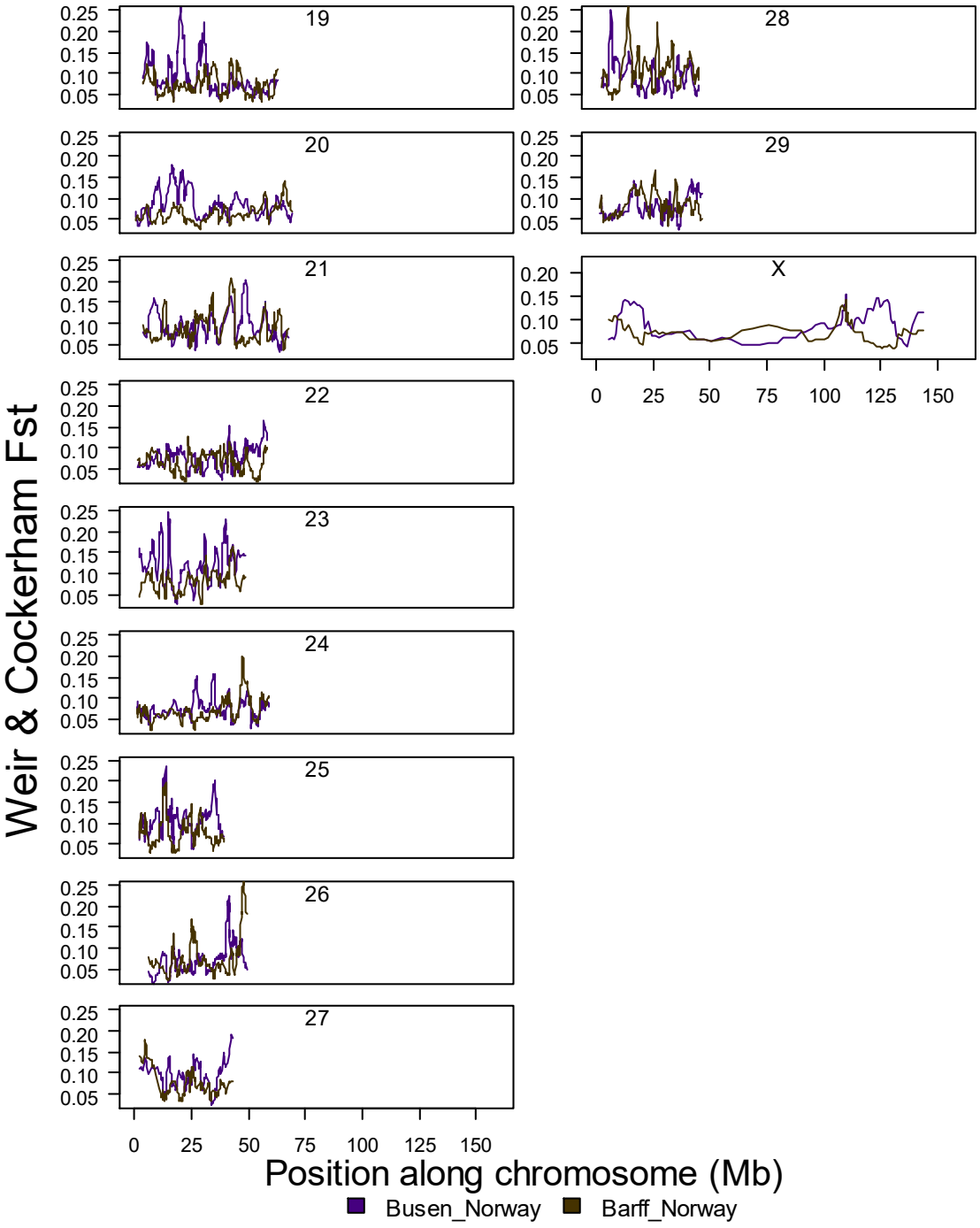


Fig.A2.9 continued.

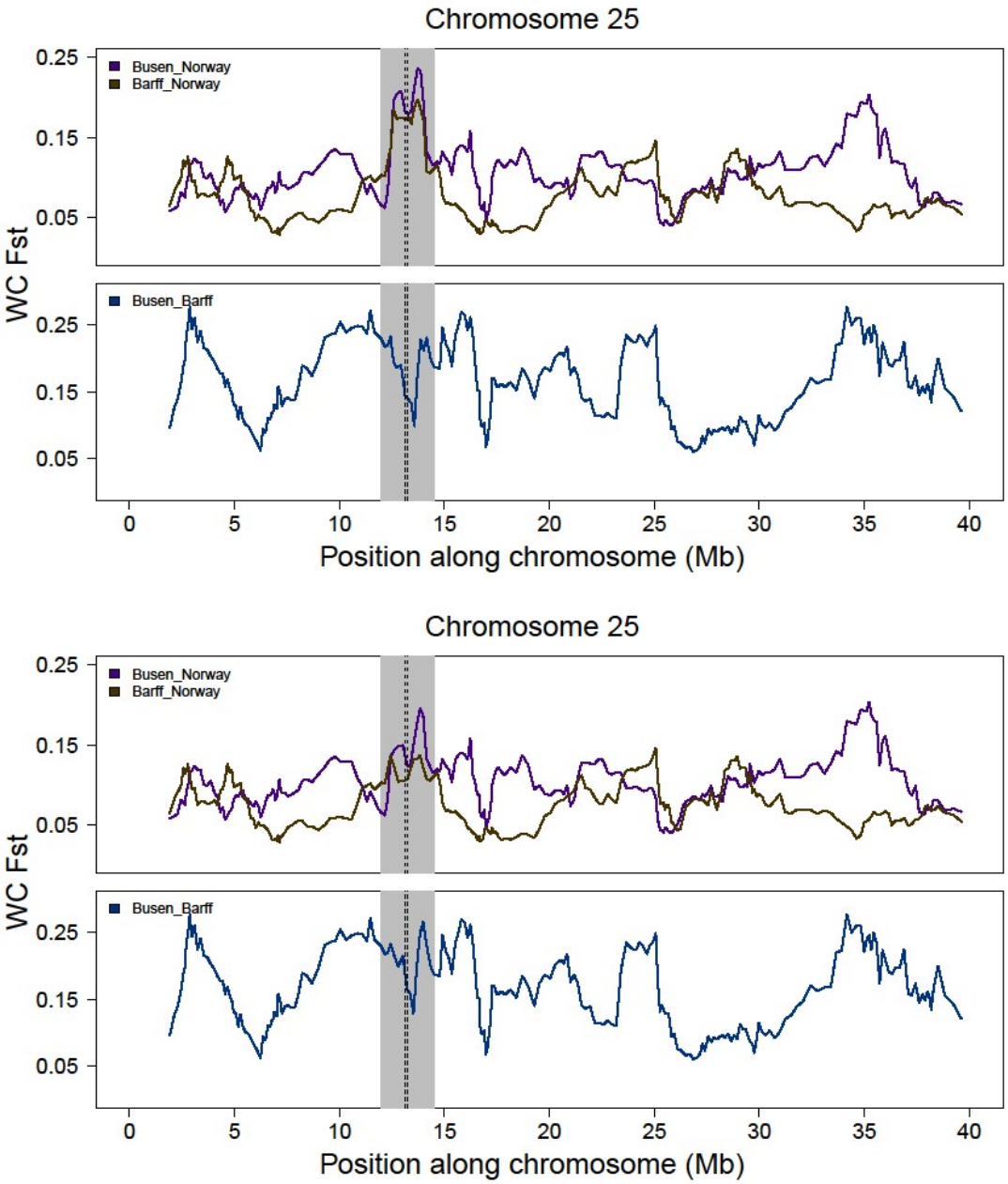


Fig. A2.10. Peak-peak-valley signal . *Sliding window Weir & Cockerham Fst plot of chromosome 25. Fst is calculated both including (above) and excluding (below) both adjacent outlier SNPs. Dotted lines indicate the positions of the two outlier SNPs. The peak-valley signal (a peak for both Busen-Norway and Barff-Norway comparison, and a valley for the Busen-Barff comparison) flattens out when removing both outliers.*

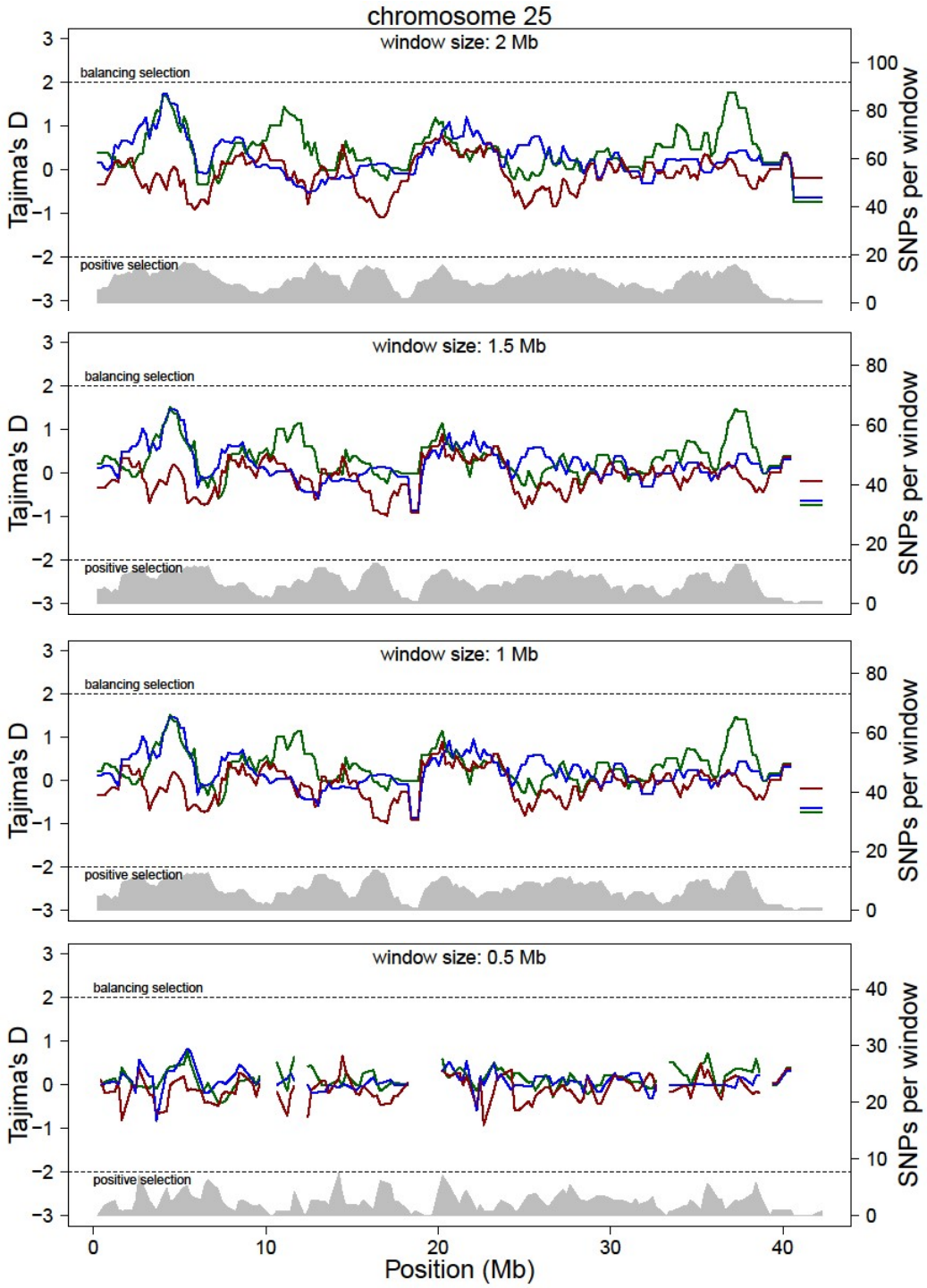


Fig.A2.11. Sliding window Tajima's D analysis. *Sliding window Tajima D scores for various window sizes (step size = 0.2 Mb) for chromosome 25 for the three study populations: Busen (blue), Barff (green), Norway (red). Grey shading indicates the number of SNPs per window.*

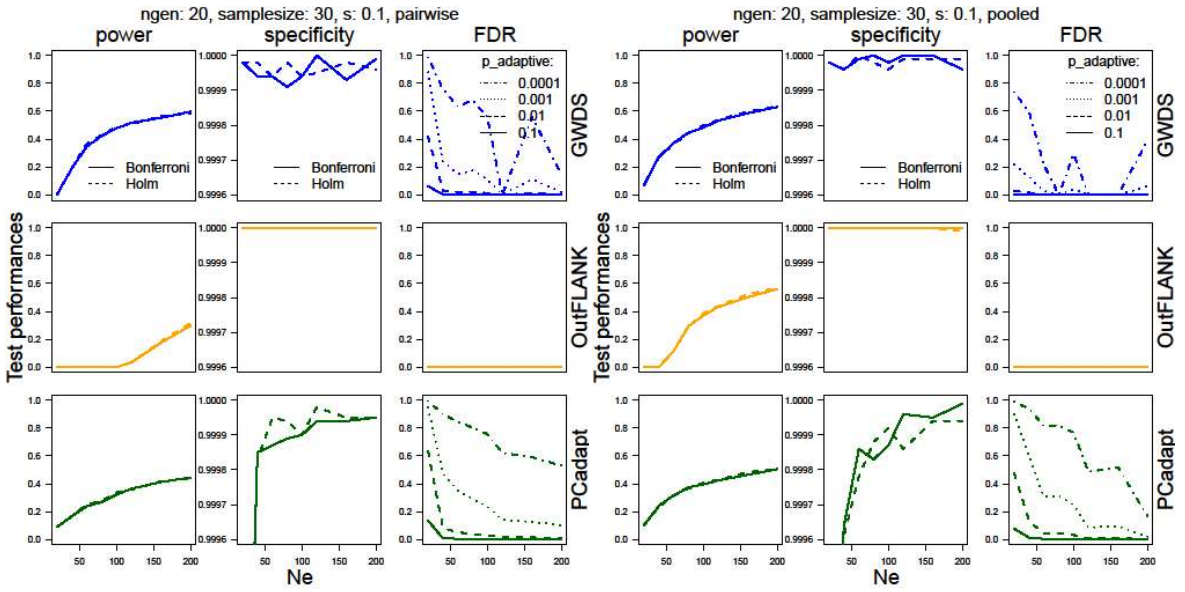


Fig. A2.12. False discovery rates (FDR) of selection scans in young founder populations. *Power, specificity and false discovery rate (FDR) estimates of the selection scans GWDS, OutFLANK and PCadapt in recently established founder populations (population age of 20 generations) given a sample size of 30 individuals per population, a selection coefficient s of strength 0.1, various constant effective population sizes (Ne) without founder bottleneck, and using either the Bonferroni or Holm multiple test correction method. Power estimates give the inverse of the false negative rate (FN), i.e. the proportion of alleles under positive selection that are correctly marked by selection scans as outliers. Specificity estimates give the inverse of the false positive rate (FP), i.e. the proportion of neutral alleles that are not marked by selection scans as outliers. The power and specificity scores are based on simulations with 90000 neutral SNPs and 10000 adaptive SNPs. FDR estimates, the proportions of false positives in the outlier set, are based on the Bonferroni power and specificity estimates and are calculated for various proportions of adaptive SNPs (p_adaptive), ranging from 10% to 0.01%, using the formula: (FP*(1-p_adaptive))/ (FP*(1-p_adaptive) + ((1-FN)*p_adaptive)). Left: pairwise approach. Right: pooled approach.*

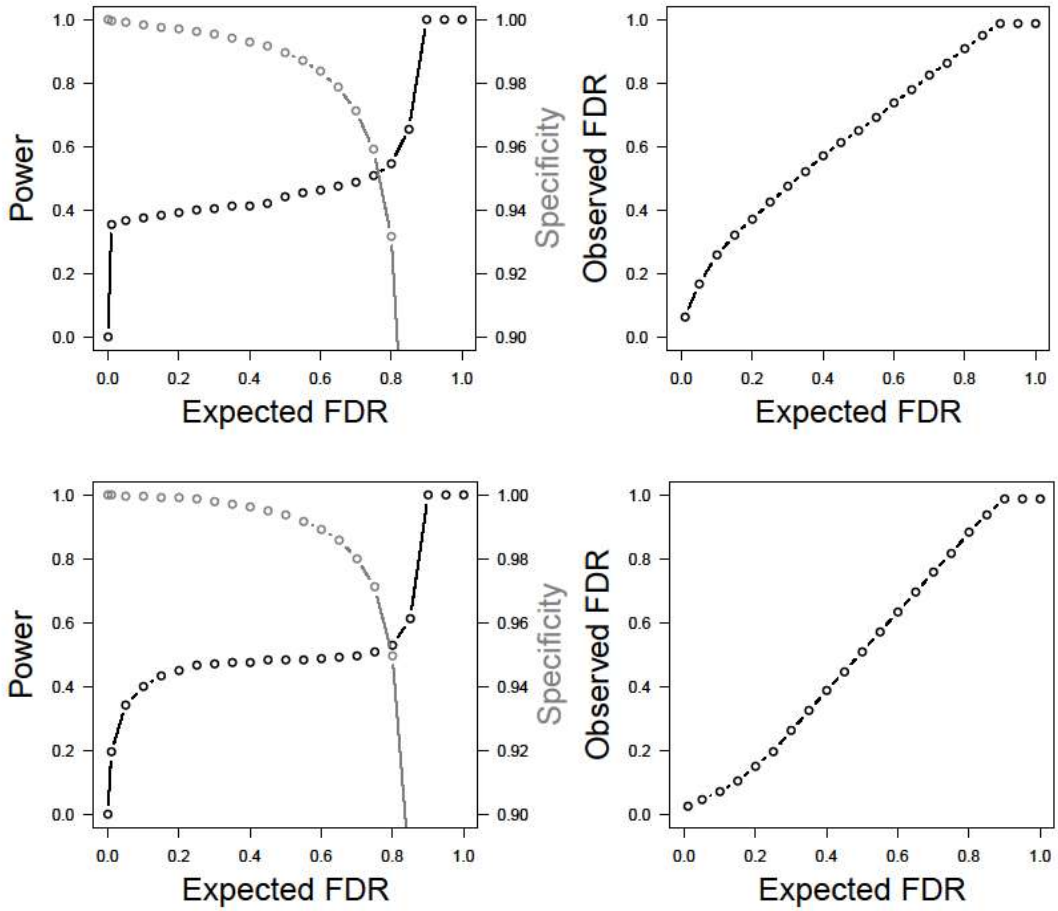


Fig. A2.13A. Bayescan power analysis. *Simulation results of Bayescan power analyses, showing power and specificity (left) and observed false discovery rates (right) in founder populations given a demographic scenario of 10 founders, a fixed Ne of 50, a population age of 20 generations. Above: pairwise approach, below: pooled approach. Analysis based on 79000 neutral and 1000 selected loci.*

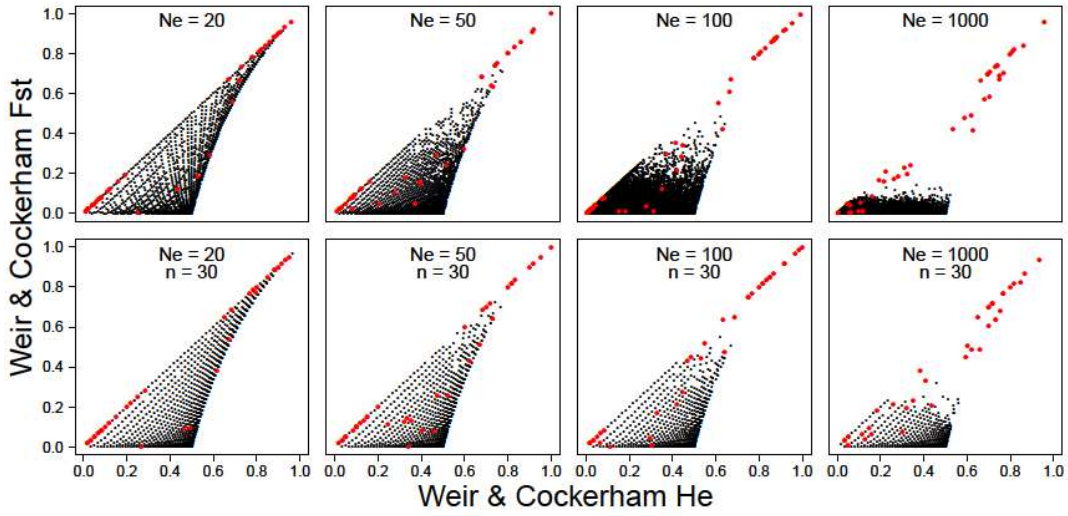


Fig.A2.14. Effect of effective population size and sampling size on Fdist distributions. *Simulated distributions of locus specific He-Fst estimates in founder populations, given a population age of 20 generations, a sample size of 30 individuals (lower row), and an uniform distribution of minor allele frequency in the source population of 0.15.Black dots are neutral SNPs; red dots are SNPs under selection (s=0.1). Number of founders equals effective population size. Red dots which are surrounded by black dots can not be detected by selection scans.*

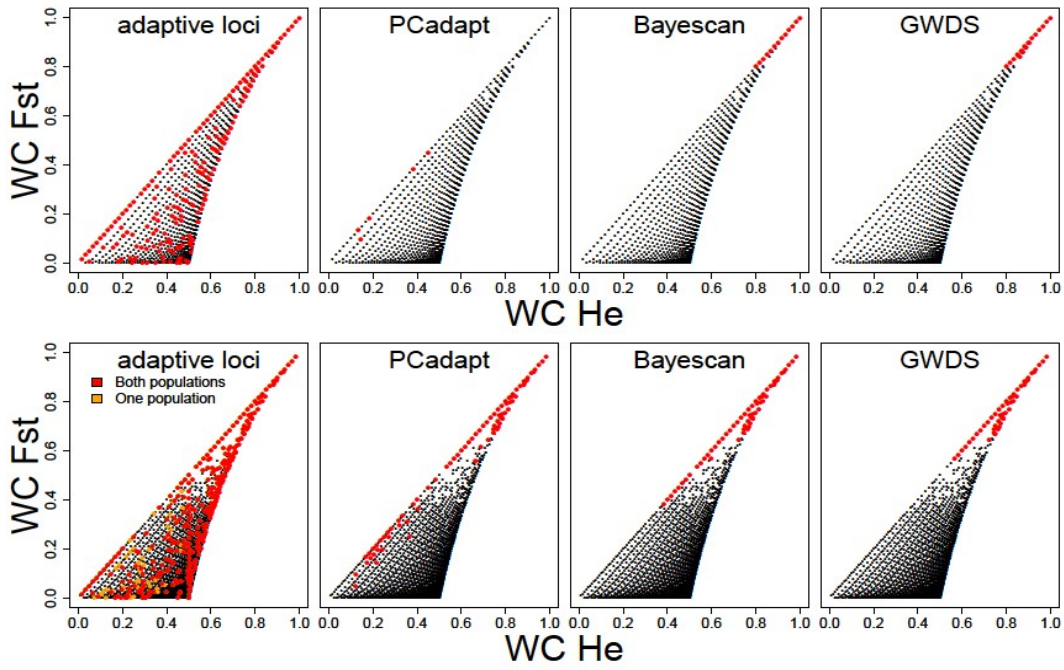


Fig.A2.15A. Detectability of outlier SNPs in pairwise versus pooled approach. *Fdist plots showing distribution of 79000 neutral (black) and 1000 selected loci (red, s=0.1) or loci marked as outlier by PCadapt, Bayescan or GWDS given a demographic scenario of 10 founders, a fixed Ne of 50, and a population age of 20 generations. Above: pairwise approach, below: pooled approach.*

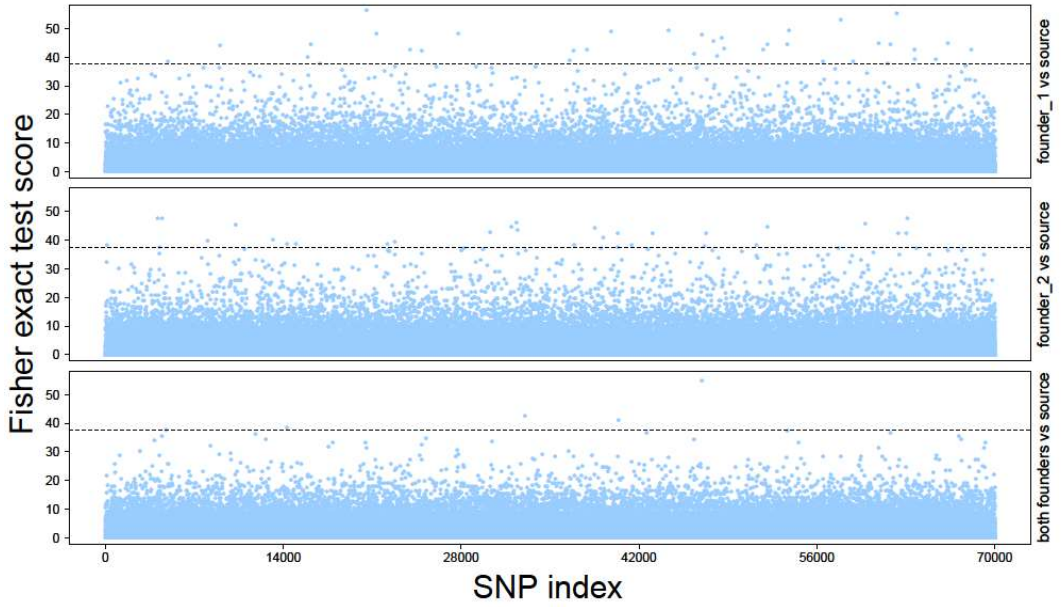
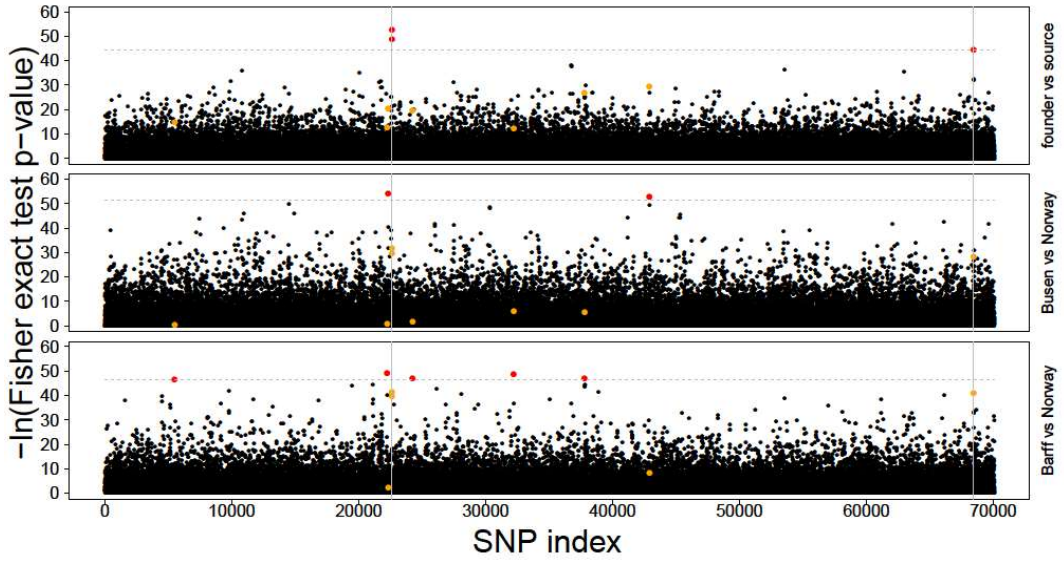


Fig.A2.15B. Simulated Fisher exact test scores test scores*. Negative log of Fisher exact test p-values on contingency tables of allele counts in simulated source and founder populations given a demographic scenario of 10 founders, a fixed Ne of 50, and a population age of 20 generations, for a dataset of 70K neutral SNPs. The ranges of the values is roughly similar to the range of values observed in the empirical dataset (see below).*



APPENDICES CHAPTER 3

Table A3.1 Sequencing output. *STACKS demultiplexing output*

Sequencing Date	ID	Total Reads	No RadTag	Low Quality	Retained	Barcode	Pool
Jul-17	G30	4720534	56228	5226	4659080	GCATG	pool1
Jul-17	G35	1757646	90512	2968	1664166	AATCG	pool1
Jul-17	G37	2140416	42458	4065	2093893	ACAGA	pool1
Jul-17	G51	4910058	63082	5864	4841112	AAGTGA	pool1
Jul-17	G54	3913612	22463	6456	3884693	ATTACA	pool1
Jul-17	G58	1726940	12046	2016	1712878	CAGGCG	pool1
Jul-17	G59	3480726	12963	3925	3463838	AGAATGA	pool1
Jul-17	8	4894792	19712	4749	4870331	AGTTAAT	pool1
Jul-17	9	2496184	7194	2960	2486030	CCACTGG	pool1
Jul-17	12	9781098	18978	10091	9752029	AGTCAAGA	pool1
Jul-17	13	8307972	20719	8284	8278969	AGTGTTAA	pool1
Jul-17	390	1025954	44599	825	980530	GCATG	pool3
Jul-17	391	2234458	62636	1467	2170355	AATCG	pool3
Jul-17	393	2071070	44902	1749	2024419	ACAGA	pool3
Jul-17	394	3946800	73517	2694	3870589	AAGTGA	pool3
Jul-17	396	3242952	25285	2872	3214795	ATTACA	pool3
Jul-17	400	1300134	10877	786	1288471	CAGGCG	pool3
Jul-17	G2	3818516	18907	3731	3795878	AGAATGA	pool3
Jul-17	G4	17348018	39769	13781	17294468	AGTTAAT	pool3
Jul-17	G8	2377182	7567	1909	2367706	CCACTGG	pool3
Jul-17	G9	4379878	14156	2966	4362756	AGTCAAGA	pool3
Jul-17	G10	4133644	12785	2726	4118133	AGTGTTAA	pool3
Jul-17	G14	3764698	78711	1978	3684009	GCATG	pool4
Jul-17	G15	2233890	78535	1032	2154323	AATCG	pool4
Jul-17	G17	3865910	70699	2499	3792712	ACAGA	pool4
Jul-17	G19	5195706	85217	2684	5107805	AAGTGA	pool4
Jul-17	G20	4666442	28656	2910	4634876	ATTACA	pool4
Jul-17	G32	1225094	16403	752	1207939	CAGGCG	pool4
Jul-17	G36	3227370	17861	1924	3207585	AGAATGA	pool4
Jul-17	G52	6919062	31703	3392	6883967	AGTTAAT	pool4
Jul-17	G53	3790266	13720	2183	3774363	CCACTGG	pool4
Jul-17	G60	3500648	21123	1829	3477696	AGTCAAGA	pool4
Jul-17	5	5628972	16518	2472	5609982	AGTGTTAA	pool4
Jul-17	10	2184836	51297	1495	2132044	GCATG	pool5
Jul-17	372	3138998	47062	2139	3089797	AATCG	pool5
Jul-17	373	2431668	34676	1927	2395065	ACAGA	pool5
Jul-17	451	9787158	57468	7354	9722336	AAGTGA	pool5
Jul-17	452	2616888	12310	1706	2602872	ATTACA	pool5
Jul-17	455	1486302	8308	883	1477111	CAGGCG	pool5
Jul-17	457	3722198	16385	2737	3703076	AGAATGA	pool5
Jul-17	M1	2535454	17111	1927	2516416	AGTTAAT	pool5

Jul-17	M3	685224	3036	596	681592	CCACTGG	pool5
Jul-17	M4	2697054	14742	2315	2679997	AGTCAAGA	pool5
Jul-17	M6	3198110	7570	2353	3188187	AGTGTTAA	pool5
Jul-17	M18	6508740	125104	10770	6372866	GCATG	pool6
Jul-17	M22	4205918	212496	6932	3986490	AATCG	pool6
Jul-17	M23	7289462	108046	9793	7171623	ACAGA	pool6
Jul-17	M26	9677048	183973	18128	9474947	AAGTGA	pool6
Jul-17	298	5211008	37274	7147	5166587	ATTACA	pool6
Jul-17	304	3788342	24326	4100	3759916	CAGGCG	pool6
Jul-17	381	10483552	35111	17618	10430823	AGAATGA	pool6
Jul-17	384	23537402	76204	36484	23424714	AGTTAAT	pool6
Jul-17	389	18578926	76254	43676	18458996	AGTCAAGA	pool6
Jul-17	392	16164504	43873	29196	16091435	AGTGTTAA	pool6
Jul-17	397	3633084	57348	1339	3574397	GCATG	pool7
Jul-17	398	3518288	54701	1414	3462173	AATCG	pool7
Jul-17	399	7367408	55452	3301	7308655	ACAGA	pool7
Jul-17	401	7327496	76341	2776	7248379	AAGTGA	pool7
Jul-17	403	7020338	23942	2859	6993537	ATTACA	pool7
Jul-17	406	1601822	10872	607	1590343	CAGGCG	pool7
Jul-17	407	3398426	15596	1470	3381360	AGAATGA	pool7
Jul-17	408	4819320	20825	1894	4796601	AGTTAAT	pool7
Jul-17	416	3125534	10657	1184	3113693	CCACTGG	pool7
Jul-17	417	2178210	10672	816	2166722	AGTCAAGA	pool7
Jul-17	418	2797654	8979	1103	2787572	AGTGTTAA	pool7
Jan-17	G5	15666388	662357	1856	15002175	GCATG	pool1
Jan-17	G11	8982408	126833	1164	8854411	AATCG	pool1
Jan-17	G12	9519226	237572	1182	9280472	ACAGA	pool1
Jan-17	G13	13534136	127336	1495	13405305	AAGTGA	pool1
Jan-17	G21	5833540	77328	689	5755523	ATTACA	pool1
Jan-17	G22	5527488	63375	625	5463488	CAGGCG	pool1
Jan-17	G23	10169118	88730	1316	10079072	AGAATGA	pool1
Jan-17	G24	13809636	162318	1773	13645545	AGTTAAT	pool1
Jan-17	G25	9087920	90647	1242	8996031	CCACTGG	pool1
Jan-17	G26	11935340	76998	1775	11856567	AGTCAAGA	pool1
Jan-17	G27	8622554	63750	1124	8557680	AGTGTTAA	pool1
Jan-17	G29	11312294	94008	1521	11216765	CACGACCA	pool1
Jan-17	G80	14785502	48283	1987	14735232	AGTTAAT	pool2
Jan-17	2	19297550	50745	2641	19244164	CCACTGG	pool2
Jan-17	4	2597432	10673	390	2586369	AGTCAAGA	pool2
Jan-17	6	15588092	36581	2177	15549334	AGTGTTAA	pool2
Jan-17	7	7761282	14653	1120	7745509	CACGACCA	pool2
Jan-17	19	9574892	33287	1248	9540357	CCACTGG	pool3
Jan-17	20	7254572	17363	974	7236235	AGTCAAGA	pool3
Jan-17	367	19445954	41878	2757	19401319	AGTGTTAA	pool3
Jan-17	368	10761344	26864	1428	10733052	CACGACCA	pool3
Jan-17	302	7453278	20169	950	7432159	CCACTGG	pool4

Jan-17	303	6721314	11194	994	6709126	AGTCAAGA	pool4
Jan-17	379	18333164	33643	2610	18296911	AGTGTTAA	pool4
Jan-17	385	20713594	40595	2854	20670145	CACGACCA	pool4
Jan-17	411	11009192	32933	1431	10974828	CCACTGG	pool5
Jan-17	412	8244678	35905	1194	8207579	AGTCAAGA	pool5
Jan-17	413	8367162	16682	1147	8349333	AGTGTTAA	pool5
Jan-17	414	7006726	21553	983	6984190	CACGACCA	pool5
Sum		643658790	5082765	386451	638189574		
% total reads		100.0	0.8	0.1	99.2		
Average		6847434	54072	4111	6789251		
Stdev		5177234	77231	6821	5151240		

Table A3.2. *Number and spacing of SNPs per chromosome based on alignment against C. elaphus (for EastAnglia, Ayrshire and Wurttemberg samples)*

chrom	**#snps**	**spacing in between adjacent snps**						
		mean	***stdv***	***min***	***0.25***	***median***	***0.75***	***max***
1	852	121384	265920	1	38	15093	132208	3771863
10	529	105298	190140	1	58	27399	144332	2003253
11	1080	128960	240830	1	61	29896	167253	3274209
12	1105	115413	190579	1	70	20694	166945	1823862
13	860	103707	182669	1	41	4793	136706	1532109
14	931	110989	194023	1	58	19159	148807	1523112
15	1092	114790	221373	1	48	15443	142108	2056297
16	504	124706	216120	1	67	26845	159649	1918948
17	487	162985	323549	1	33	4142	199111	2881002
18	1098	138719	231852	1	65	30552	190461	2047684
19	935	134826	238129	1	57	23612	171611	1977042
2	619	101185	202223	1	47	6253	125468	2201572
20	1319	112982	241913	1	43	8848	134445	4275578
21	849	126312	224512	1	50	16138	161642	2067827
22	567	112661	198039	1	51	20995	143018	1350892
23	1187	90833	182761	1	43	6532	116395	2875907
24	736	106254	169630	1	51	17424	159048	1281853
25	756	127430	254696	1	37	11043	161197	3444338
26	496	110888	192304	1	70	23062	166248	1713723
27	789	106783	193455	1	50	12716	143034	1898917
28	584	140090	293676	1	42	3525	180974	3082616
29	637	125176	233127	1	42	12070	146790	2061521
3	629	139457	241912	1	49	20471	175374	1554935
30	849	137923	328622	1	49	21188	148921	4677516
31	373	201978	353675	1	54	41326	261611	1845100
32	502	118840	222076	1	51	8982	143224	1545533
33	777	155295	274170	1	59	24351	186770	1847286
4	844	95416	171612	1	45	4310	126667	1075264
5	1670	105966	206628	1	53	11200	135567	2803155
6	498	143531	271524	1	44	4749	187056	2452052
7	661	98527	187743	1	50	13370	129682	1880328
8	475	117420	195229	1	74	23230	162187	1094329
9	1179	120283	228460	1	49	15083	138311	2718935
X	500	362326	862053	1	26	400	363220	7751744
Y	10	40	74	1	1	17	28	234
contigs	319							
mean	***793***	***129980***	***247801***	***1***	***51***	***16026***	***163413***	***2420891***
stdev	***293***	***46462***	***117682***	***0***	***11***	***9416***	***44755***	***1279768***

Table A3.3. SNP dataset summary statistics for main dataset (above) and Aurignac dataset (below)

	Before filtering	**After filtering**	**After thinning**
Number of individuals	94	78	78
Number of SNPs	52364	31459	15697
Percentage of SNPs with maf >= 0.05	57.9	68.39	66.64
Mean spacing between SNPs	43897.56	43686.37	89216.5
Median spacing between SNPs	368	369	54451
Mean proportion of missing data per individual	0.15	0.07	0.07
GC content	0.62	0.62	0.63
Transition vs transversion ratio	2.43	2.59	2.66

	Before filtering	**After filtering**	**After thinning**
Number of individuals	30	29	29
Number of SNPs	29488	19992	10732
Percentage of SNPs with maf >= 0.05	80.96	91.53	91.4
Mean spacing between SNPs	75241.85	78510.37	150582.83
Median spacing between SNPs	1156.5	5017	91906
Mean proportion of missing data per individual	0.06	0.05	0.05
GC content	0.59	0.6	0.61
Transition vs transversion ratio	2.44	2.66	2.77

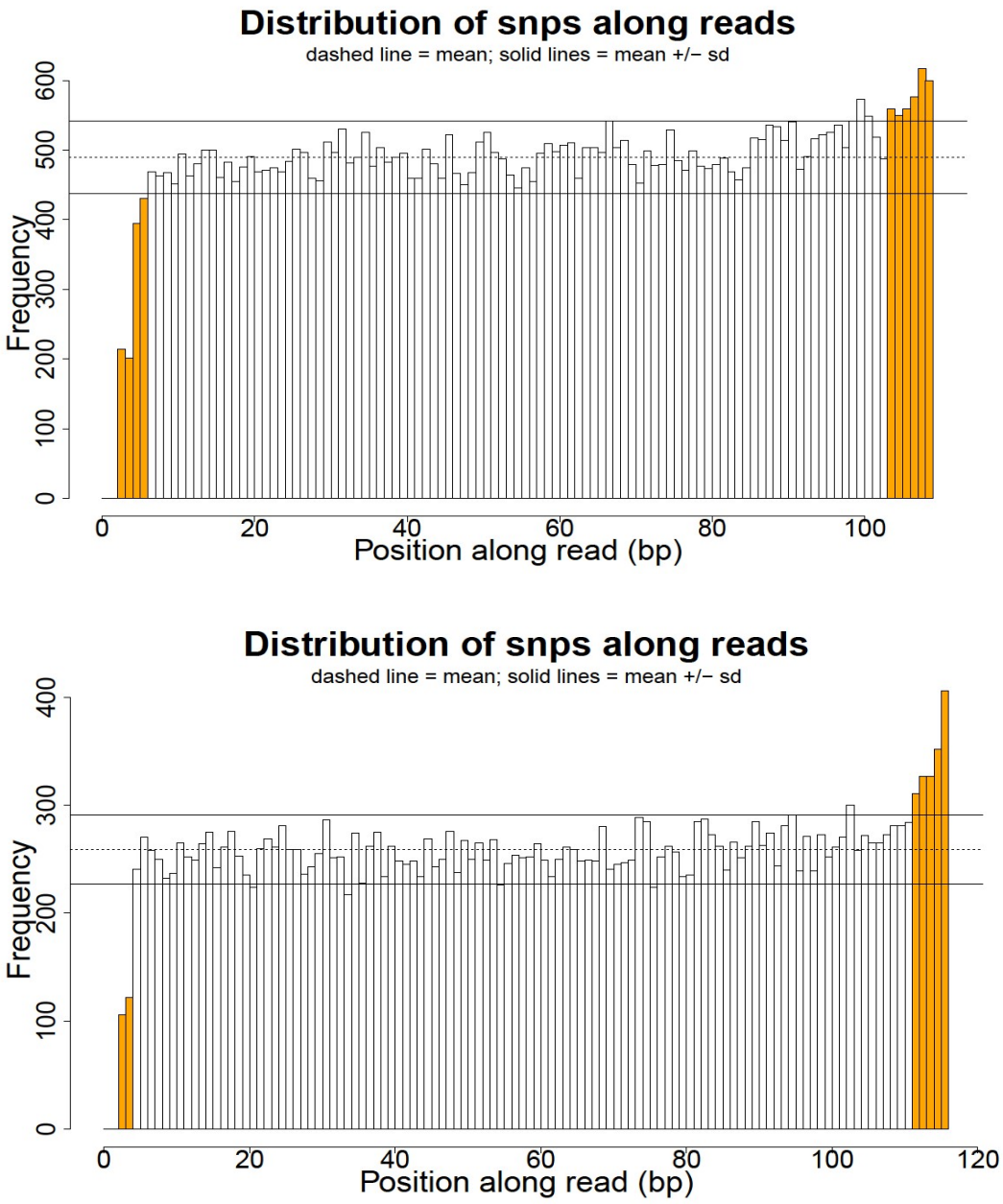


Fig. A3.1. Distribution of SNPs along sequencing reads. *Upper. Ayrshire, EastAnglia and Wurttemberg dataset. Lower: Aurignac dataset.*

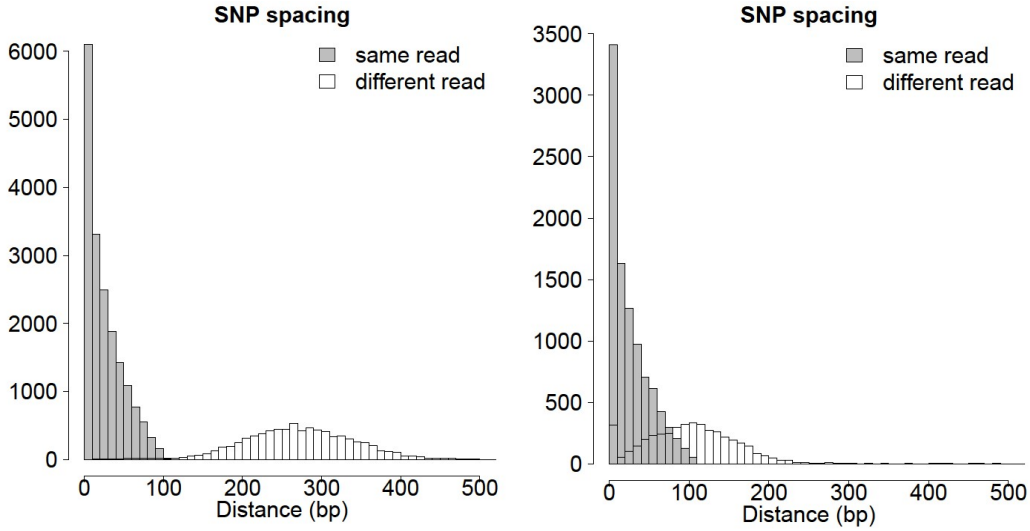


Fig. A3.2. SNP spacing. *Left: Ayrshire, EastAnglia and Wurttemberg dataset. Right: Aurignac dataset.*

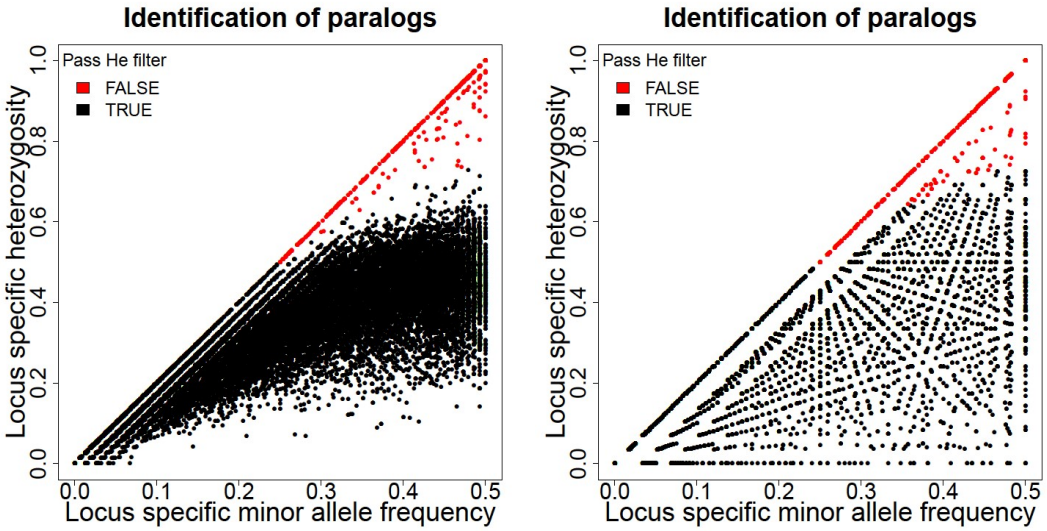


Fig. A3.3. Distribution of SNPs along sequencing reads. *Left: Ayrshire, EastAnglia and Wurttemberg dataset. Right: Aurignac dataset.*

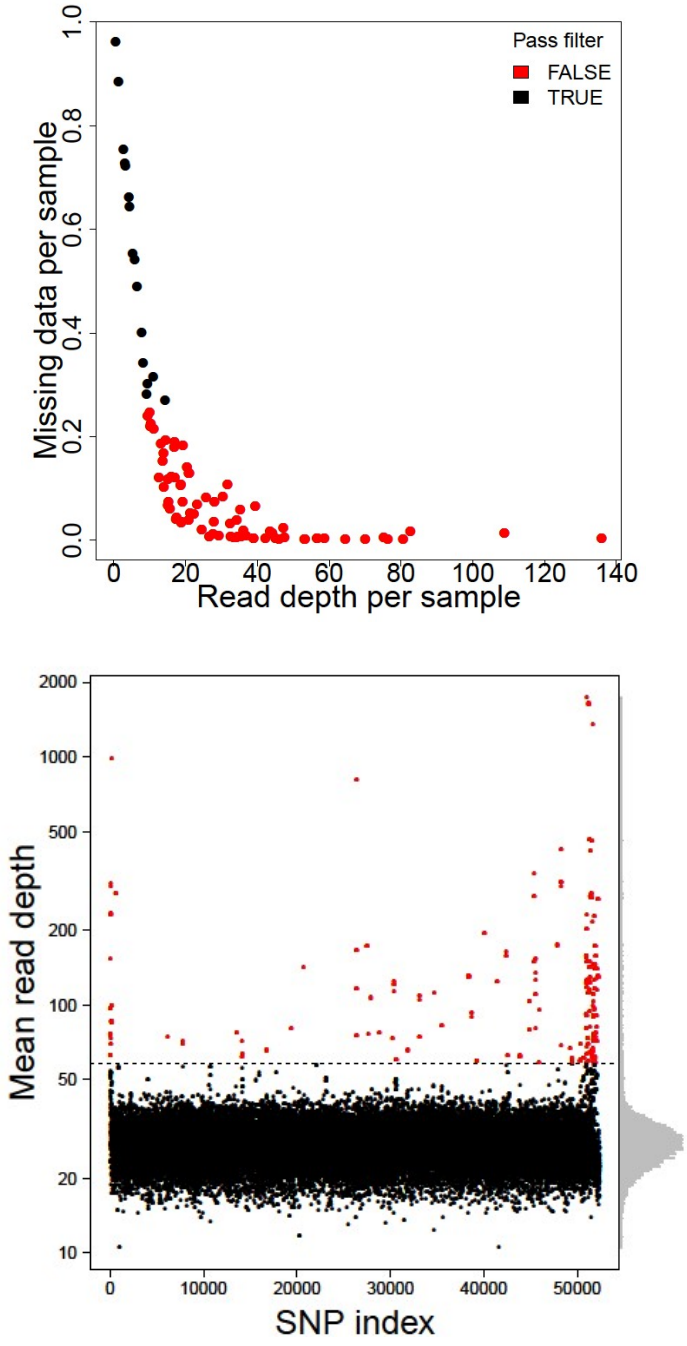


Fig. A3.4. Read depth per sample and per locus.

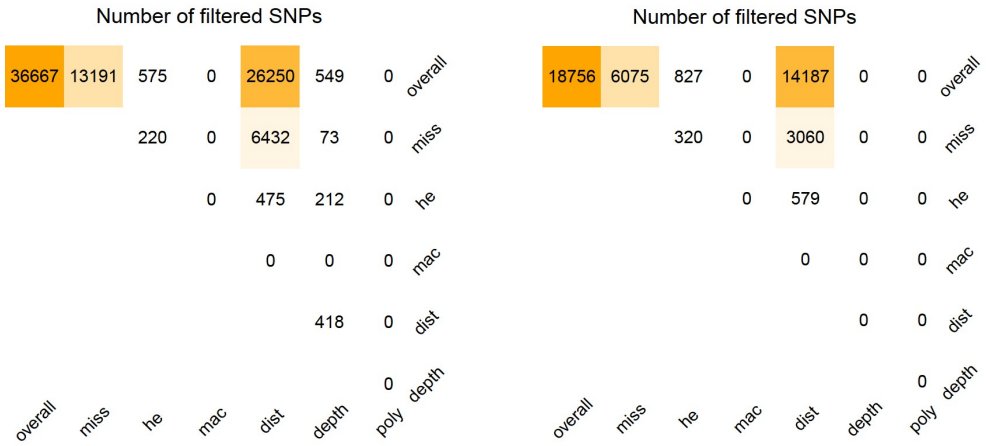


Fig. A3.5. Filter statistics. *Left: Ayrshire, EastAnglia and Wurttemberg dataset. Right: Aurignac dataset.*

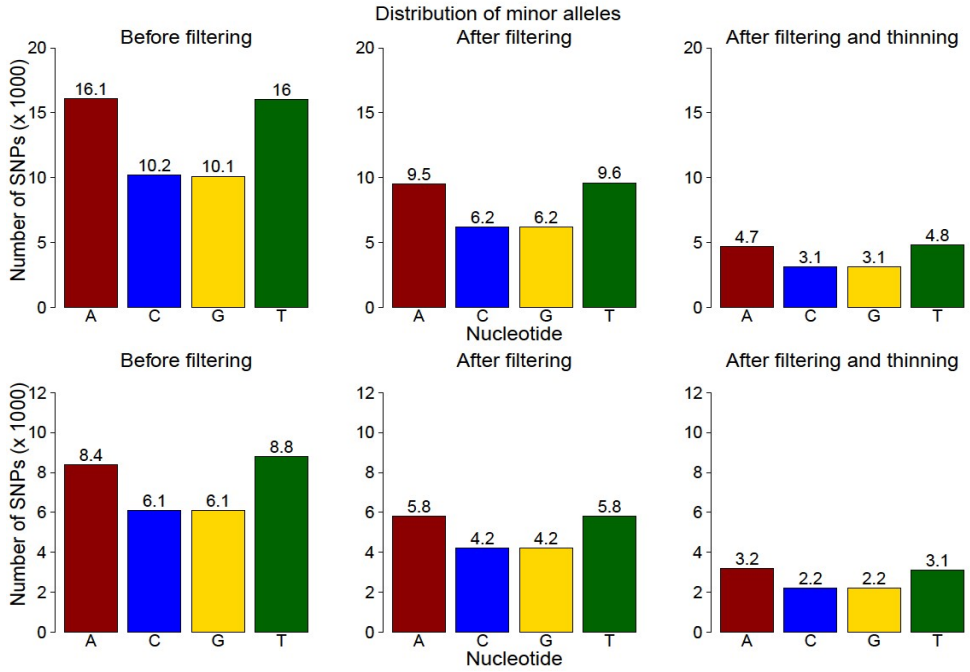


Fig. A3.6. GC-ratio.

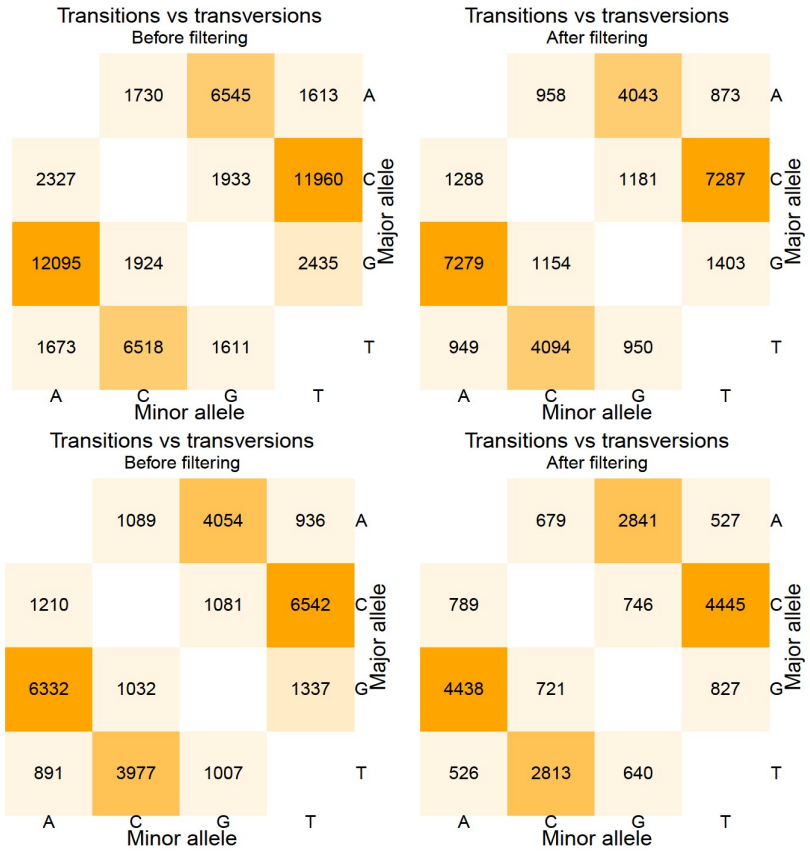


Fig. A3.7. Transition vs transversion ratio.

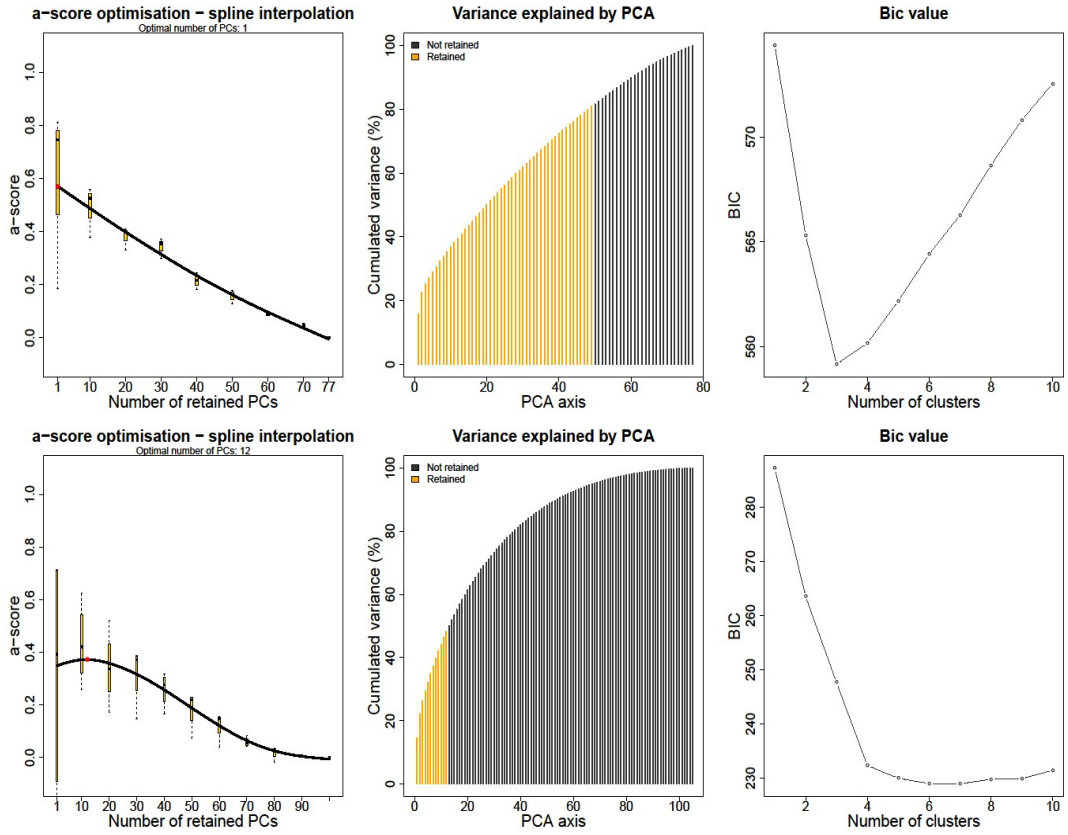


Fig. A3.8A. DAPC summary statistics. *Above: Ayrshire, East Anglia and Wurttemberg (AEW) dataset. Below: intersect datasect, which consists of SNPs occuring in both the AEW and the Aurignac dataset.*

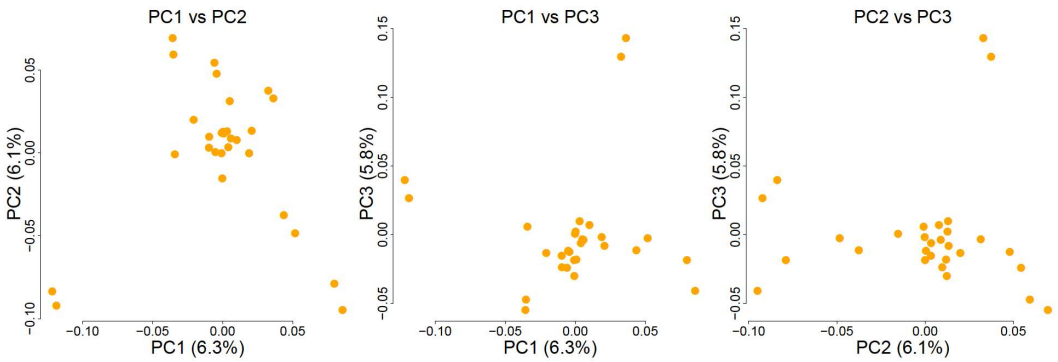


Fig. A3.8B. PCoA analyses Aurignac dataset. *PCoA analyses based on Hamming's genetic distance and based on a dataset of 10K SNPs, suggest absence of population structure in the Aurignac dataset, except for the 2 samples which appear relatively unrelated.*

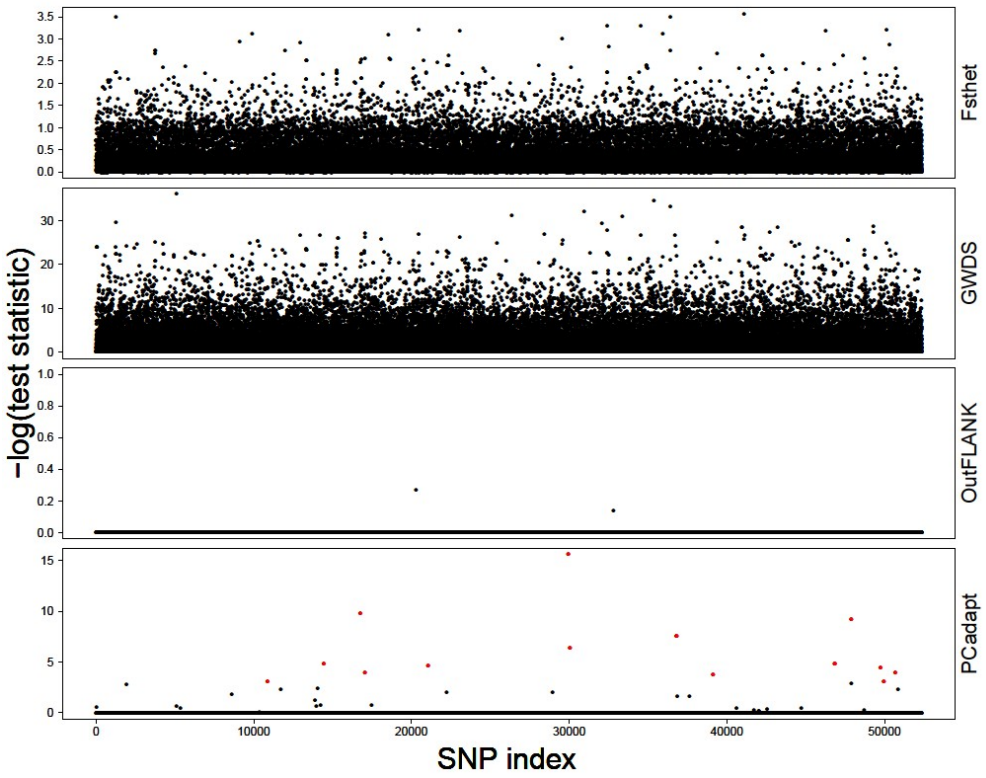


Fig. A3.9A. Selection scan test scores for modern UK vs modern mainland. *Negative log of selection scans (i.e. Fsthet, GWDS, OutFLANK and PCadapt) for both modern day UK populations (i.e. samples from Ayrshire and EastAnglia combined) vs Germany.*

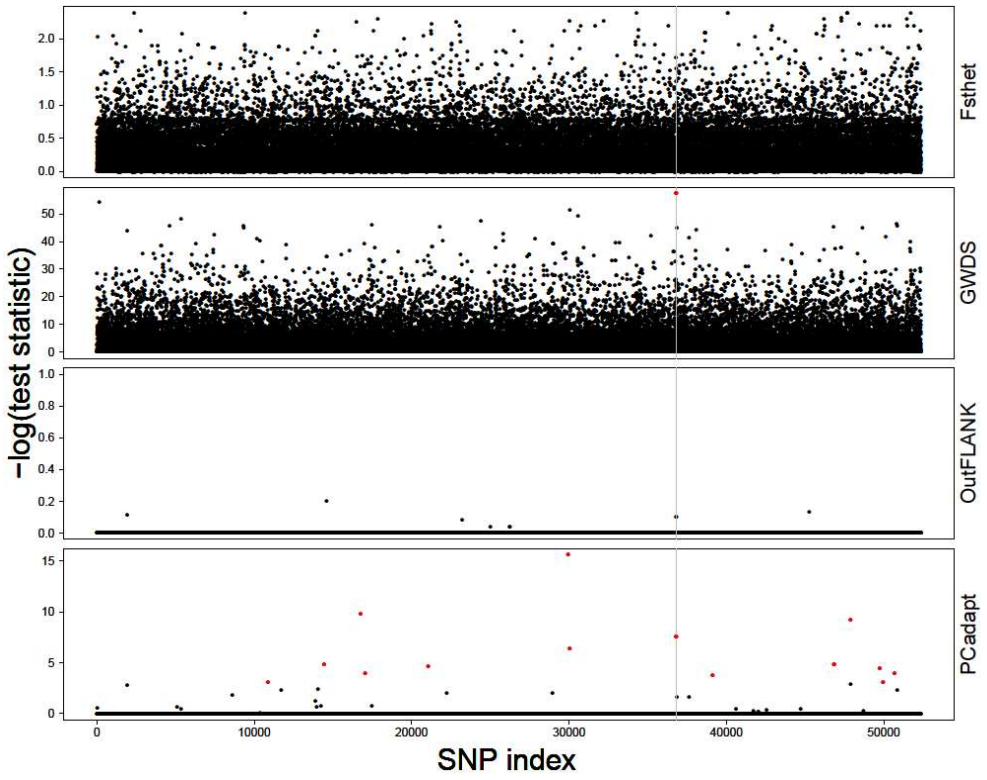


Fig. A3.9B. Selection scan test scores for native UK vs native mainland. *Negative log of selection scans (i.e. Fsthet, GWDS, OutFLANK and PCadapt) for the native UK populations (i.e. Ayrshire) vs native mainland populations (i.e. samples from EastAnglia and Germany combined).*

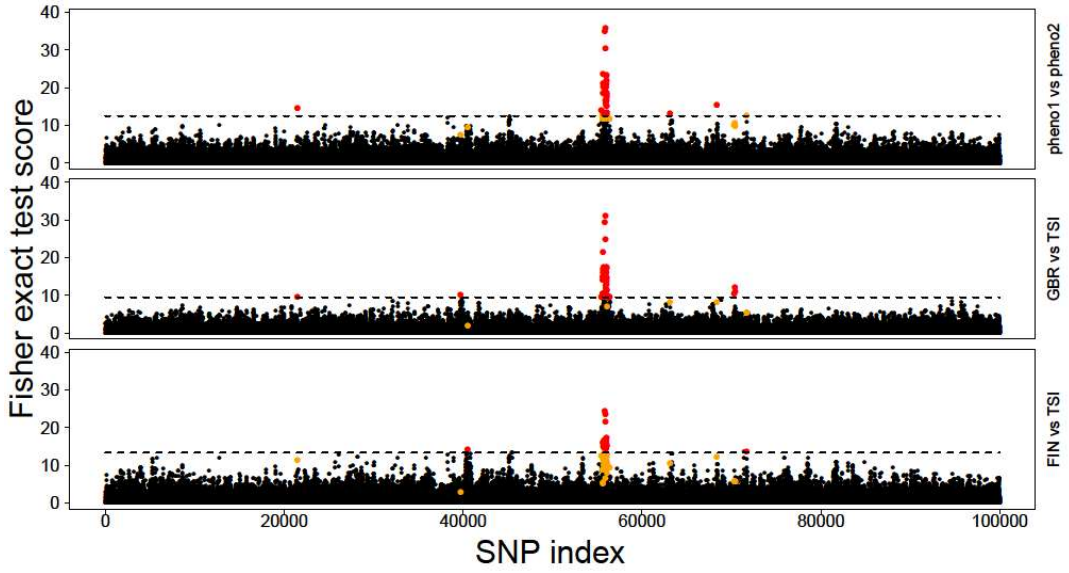


Fig. A3.10A. GWDS test results for the control IGSR human dataset (chromosome 2) highlighting outlier region associated with lactose tolerance. *GBR = Great-Britain (lac+), FIN = Finland (lac+), TSI = Toscane (lac-). Shown are the negative log of Fisher exact test-p-values on allele count tables. An outlier region is detected for the pooled comparison (GBR and FIN combined vs TSI) as well as for both pairwise comparisons (GBR vs TSI and FIN vs TSI). Red: SNPs marked as outlier for the actual comparison. Orange; SNPs marked as outlier in another comparison.*

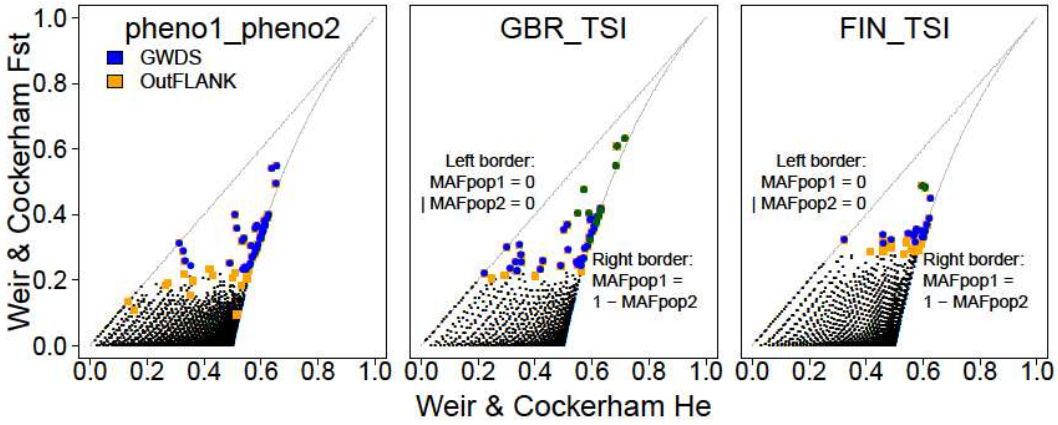


Fig. A3.10B. Fdist plots showing selection scan results for the control IGSR human dataset (chromosome 2). *GBR = Great-Britain (lac+), FIN = Finland (lac+), TSI = Toscane (lac-). Shown are locus specific Weir and Cockerham 1987 heterozygosity and Fst estimates, for the pooled comparison (pheno1: GBR and FIN combined; pheno2: TSI) as well as for both pairwise comparisons (GBR vs TSI and FIN vs TSI). Blue, orange and green dots indicate SNPs marked as outliers by respectively GWDS, OutFLANK and Pcadapt.*

APPENDICES CHAPTER 4

Table A4.1.Genome wide heterozygosity. *Number and percentages of observed heterozygosite sites after mapping sequencing reads against reference genomes for European roe deer (C. capreolus) and Siberian roe deer (C. pygargus). White tailed deer (O. virginianus) and red deer (C. elaphus) were included for comparison.*

	C. capreolus	**C. pygargus**	**C. pygargus *down-sampled***	**O. virginianus**	**Cervus elaphus**
all read depths					
heterozygous sites	3,539,884	7,787,918			
total sites	2,444,157,882	2,511,732,823			
heterozygosity	0.140%	0.310%			
read depth ≥ 8					
mean read depth	21.1	39	21.1		
heterozygous sites	3,119,328	7,711,705	6,627,891	10,268,498	2,641,723
total sites	2,177,801,796	2,409,308,248	2,230,619,899	2,075,403,030	1,876,495,191
heterozygosity	0.143%	0.320%	0.297%	0.495%	0.141%

Table A4.2.Single nucleotide variations (SNVs). *Number of fixed and segregating single nucleotide variations (SNVs) between C. pygargus and C. capreolus, inferred from crossmapping raw reads to the reference genome of the sister species. Sequence dissimilarity is calculated as: (0.5*segregating SNVs + fixed SNVs)/total sites*100.*

raw reads	**C. pygargus**		**C. capreolus**	
reference genome	**C. capreolus**		**C. pygargus**	
segregating SNVs	6,649,612	0.324	3,291,379	0.156
fixed SNVs	8,997,247	0.438	13,047,514	0.618
fixed transitions	6,299,219	70.0	9,364,731	71.8
fixed transversions	1,130,784	12.6	1,691,780	13.0
fixed ambiguous	1,467,244	16.3	1,991,003	15.3
total sites	2,051,852,399		2,112,183,935	
sequence dissimilarity (%)		0.600		0.696

Table A4.3A. PAML's codeML likelihood ratio test (LRT) scores of genes marked by codeml as under positive selection in the genus Capreolus. *Ln0: log likelihood of neutral model (i.e. all codons: ω ≤ 1). Ln1: log likelihood of nested model with positive selection (i.e. some codons: ω > 1). Np0: number of parameters of neutral model. Np1: number of parameters of positive selection model. D-statistic: 2(ln1 – ln0). p-value: Chi-squared test p-value associated with D-statistics and 1 degree of freedom (np1 – np0). logp: negative log10 of p-value.*

gene	ln0	ln1	np0	np1	D-statistic	p-value	-logp
g02979.t1	-6733.63	-6722.26	30	31	22.72307	1.87E-06	5.727907
g03242.t1	-2076.81	-2061.09	30	31	31.44965	2.05E-08	7.688925
g03905.t1	-4794.54	-4782.87	30	31	23.33693	1.36E-06	5.866574
g05086.t1	-1193.22	-1171.07	30	31	44.30169	2.81E-11	10.55056
g06421.t1	-4318.72	-4286.79	30	31	63.85284	1.33E-15	14.87541
g06831.t1	-2070.1	-2054.18	30	31	31.85099	1.66E-08	7.778677
g06841.t1	-4929.99	-4910.34	30	31	39.29744	3.64E-10	9.438997
g07795.t1	-1606.52	-1586.01	30	31	41.00313	1.52E-10	9.818198
g11691.t1	-38130	-38110.8	30	31	38.3211	6.00E-10	9.221774
g12098.t1	-8749.9	-8731.18	30	31	37.4581	9.34E-10	9.029663
g13077.t1	-1646.77	-1635.37	30	31	22.79416	1.80E-06	5.743973
g13226.t1	-5803.29	-5786.23	30	31	34.13433	5.14E-09	8.288734
g16296.t1	-1425.03	-1411.3	30	31	27.45362	1.61E-07	6.793424
g17080.t1	-796.723	-784.594	30	31	24.25775	8.43E-07	6.074341
g17279.t1	-3280.78	-3265.08	30	31	31.39737	2.10E-08	7.677232
g20881.t1	-4595.25	-4576.94	30	31	36.60754	1.44E-09	8.840216
g20973.t1	-13263	-13114.9	30	31	296.1882	0	Inf
g21474.t1	-2422.03	-2405.53	30	31	33	9.22E-09	8.035462

Table A4.3C. Names and characteristics of genes marked by codeml as under positive selection in the genus Capreolus. *Gene codes and names are inferred using online blast tool of ncbi webpage. Potential outlier column indicates presence of 1 or more lineage specific amino acid mutations (LSAAM) with a BEB-score above 0.5 for class2a or class2b.*

gene	code	name	Potential outlier	visual_check
g02979.t1	ARHGAP33	Rho GTPase activating protein 33	TRUE	5 LSAAM with BEB>0.5, of which 1 due to missing data in *C. capreolus*; the other 4 credible
g03242.t1	RAB22 1	member RAS oncogene family	FALSE	No data for H. inermus; codon 5 in *C. capreolus* is stopcodon; sequence is completely different from *C. pygargus*
g03905.t1	NLK	nemo like kinase	TRUE	5 (possibly 6) LSAAM, two of which adjacent.
g05086.t1	SGO1	shugoshin 1	FALSE	This is a dubious one: incomplete data, and most LSAAMs with BEB>0.5 due to misalignment; 3 LSAAMare credible though, no clustering
g06421.t1	DGKA	diacylglycerol kinase alpha	FALSE	2 LSAAM, plus region with 16 adjacent LSAAM, but with missing data for 9 out of 15 species
g06831.t1	RAB22A	member RAS oncogene family	FALSE	paralog comparison, *C. capreolus* as a result completely different from *C. pygargus*
g06841.t1	PAXBP1	PAX3 and PAX7 binding protein	TRUE	1 LSAAM with BEB>0.5, plus potentially 12 close to each other, but can not be confirmed due to missing data in *C. capreolus*
g07795.t1	unknown	unknown	FALSE	5 LSAAM at end of gene with BEB>0.5, but due to misalignment
g11691.t1	MDN1	midasin AAA ATPase	TRUE	Definite candidate for gene under positive selection in genus. Over 30 LSAAM (of which 2 with BEB>0.5) spread throughout gene of 16 kb with many mutations (hence not found be accelerated dN/dS test). A proportion of LSAAM's are shared with distant clades. The low number of LSAAM with BEB>0.5 might be due to paralog comparison in C. elaphus and B. taurus.
g12098.t1	ADGRB1	adhesion G protein-coupled receptor B1	FALSE	1 LSAAM with BEB>0.5 in end of gene, but in region with no data for *C. capreolus*

g13077.t1	no_hits	no_hits	TRUE	5 adjacent LSAAM with BEB>0.5, mutation of A-repeat into G-repeat
g13226.t1	KLHL29	kelch like family member 29	TRUE	cluster of 5 LSAAM in 6 codons, 4 with BEB>0.5 in class2b and 1 with BEB in class2b, in a variable region with putative paralogs, but none paralog with same sequence as *C. capreolus* and *C. pygargus*
g16296.t1	unknown	unknown	FALSE	high proportion of missing data, no BEB>0.5
g17080.t1	BOLA	histocompatibility antigen alpha chain BL3-7	TRUE	cluster of 4 adjacent LSAAM with BEB>0.5, but missing data in many other species
g17279.t1	TOP1	topoisomerase	FALSE	Most likely H. inermus stands out rather than *Capreolus*, but due to missing data in other species difficult to tell
g20881.t1	?	PRAME family member 9	FALSE	paralog comparison, *C. capreolus* as a result very similar to H. inermus and O. hemionus, and very different from *C. pygargus*
g20973.t1	?	lysine specific demethylase 6A	FALSE	paralog comparison
g21474.t1	ZCCHC18	zinc finger CCHC domain containing 18	TRUE	paralogs present, but does not affect the interpretation of three LSAAM with BEB>0.95

Table A4.4A. PAML's codeML likelihood ratio test (LRT) scores of genes marked by codeml as under positive selection in the species C. capreolus. *Ln0: log likelihood of neutral model (i.e. all codons: ω ≤ 1). Ln1: log likelihood of nested model with positive selection (i.e. some codons: ω > 1). Np0: number of parameters of neutral model. Np1: number of parameters of positive selection model. D-statistic: 2(ln1 – ln0). p-value: Chi-squared test p-value associated with D-statistics and 1 degree of freedom (np1 – np0). logp: negative log10 of p-value.*

gene	ln0	ln1	np0	np1	D	pvalue	logp
g00068.t1	-2734.05	-2721.13	30	31	25.84505	3.70E-07	6.431854
g00884.t1	-768.806	-743.406	30	31	50.79874	1.02E-12	11.98995
g01006.t1	-3258.85	-3245.06	30	31	27.59647	1.49E-07	6.825501
g01347.t1	-2432.39	-2409.79	30	31	45.1868	1.79E-11	10.74689
g01369.t1	-1662.24	-1641.09	30	31	42.30389	7.81E-11	10.10714
g01907.t1	-2600.16	-2574.99	30	31	50.34608	1.29E-12	11.88979
g01951.t1	-1101.68	-1084.06	30	31	35.23819	2.92E-09	8.534994
g01978.t1	-3020.26	-3000.17	30	31	40.17492	2.32E-10	9.634118
g02746.t1	-2278.56	-2260.15	30	31	36.81401	1.30E-09	8.886214
g03119.t1	-6660.71	-6644.92	30	31	31.57501	1.92E-08	7.716964
g04152.t1	-15391.1	-15361.2	30	31	59.69658	1.11E-14	13.95459
g04403.t1	-2732.37	-2718.63	30	31	27.47576	1.59E-07	6.798396
g04906.t1	-5710.4	-5695.3	30	31	30.20103	3.90E-08	7.409489
g05061.t1	-28065.1	-28045.6	30	31	38.98903	4.26E-10	9.370392
g05187.t1	-2653.51	-2637.96	30	31	31.09114	2.46E-08	7.608724
g05450.t1	-1996.89	-1978.04	30	31	37.70009	8.25E-10	9.083541
g05700.t1	-7486.2	-7406.72	30	31	158.9567	0	Inf
g06407.t1	-3319.86	-3298.59	30	31	42.52834	6.97E-11	10.15698
g06580.t1	-2481.65	-2468.36	30	31	26.58247	2.53E-07	6.597697
g06951.t1	-4092.69	-4080.52	30	31	24.32439	8.14E-07	6.089365
g07049.t1	-2996.6	-2975.01	30	31	43.17997	4.99E-11	10.30165
g07071.t1	-6874.24	-6851.7	30	31	45.08549	1.89E-11	10.72442
g07241.t1	-2123.87	-2110.74	30	31	26.26559	2.98E-07	6.526449
g07440.t1	-4906.12	-4894.94	30	31	22.36866	2.25E-06	5.647786
g07609.t1	-2549.05	-2537.58	30	31	22.93205	1.68E-06	5.775131
g07689.t1	-4134.23	-4121.45	30	31	25.57091	4.26E-07	6.370162
g07894.t1	-3013.14	-2999.96	30	31	26.3703	2.82E-07	6.549996
g08423.t1	-4735.8	-4720.63	30	31	30.33946	3.63E-08	7.440486
g08732.t1	-10032	-10016.6	30	31	30.91369	2.70E-08	7.569019
g09357.t1	-10552.1	-10525.5	30	31	53.10246	3.17E-13	12.49944
g09440.t1	-2662.23	-2644.22	30	31	36.02477	1.95E-09	8.710355
g09476.t1	-16954.5	-16935.5	30	31	37.89371	7.47E-10	9.126647
g10392.t1	-12417	-12379.2	30	31	75.49133	0	Inf
g11268.t1	-2352.5	-2326.77	30	31	51.47516	7.25E-13	12.13961
g11598.t1	-3712.08	-3700.15	30	31	23.84842	1.04E-06	5.982019
g11691.t1	-38126.4	-38114.8	30	31	23.15561	1.49E-06	5.825631
g11849.t1	-770.484	-754.55	30	31	31.86789	1.65E-08	7.782457
g12015.t1	-1384.84	-1367.13	30	31	35.43809	2.63E-09	8.579571
g12222.t1	-12357.7	-12338.5	30	31	38.35916	5.89E-10	9.230245
g12484.t1	-1703.72	-1682.05	30	31	43.3365	4.61E-11	10.33639
g12563.t1	-4452.8	-4417.27	30	31	71.06437	0	Inf
g12719.t2	-16482.9	-16470.7	30	31	24.33606	8.09E-07	6.091996
g12882.t1	-5738.79	-5721.35	30	31	34.89088	3.49E-09	8.457534
g13700.t1	-5036.86	-5009.28	30	31	55.17739	1.10E-13	12.95808
g13760.t1	-3485.35	-3472.45	30	31	25.78503	3.82E-07	6.41835
g13957.t1	-3562.58	-3542.02	30	31	41.11381	1.44E-10	9.842791
g14096.t1	-1160.5	-1149	30	31	22.99178	1.63E-06	5.788624
g14583.t1	-1706.58	-1675.72	30	31	61.72818	4.00E-15	14.39829
g15865.t1	-2558.72	-2538.94	30	31	39.54982	3.20E-10	9.495126
g16824.t1	-15622.2	-15607.2	30	31	30.01032	4.30E-08	7.366781
g17856.t1	-2344.38	-2319.82	30	31	49.11296	2.42E-12	11.61683
g18061.t1	-957.502	-943.288	30	31	28.4278	9.73E-08	7.012073
g18092.t1	-1766.6	-1753.2	30	31	26.798	2.26E-07	6.646139
g18276.t1	-6440.45	-6425.12	30	31	30.64572	3.10E-08	7.509046
g18672.t1	-7095.9	-7077.26	30	31	37.27622	1.03E-09	8.98916
g18675.t1	-12112.1	-12094.6	30	31	34.9216	3.43E-09	8.464388
g18676.t1	-11979.2	-11962	30	31	34.24685	4.85E-09	8.313847
g18726.t1	-39105.6	-39071.2	30	31	68.96616	1.11E-16	15.95459
g18830.t2	-9656.81	-9625.52	30	31	62.58891	2.55E-15	14.59286

g18911.t1	-3315.97	-3284.18	30	31	63.58698	1.55E-15	14.80846
g19145.t1	-3188.36	-3171.61	30	31	33.49442	7.15E-09	8.145883
g19891.t1	-16015.4	-15975.6	30	31	79.67064	0	Inf
g21058.t1	-1071.7	-1056.82	30	31	29.7715	4.86E-08	7.313287
g21142.t1	-3181.88	-3165.5	30	31	32.77273	1.04E-08	7.984691
g21241.t1	-12705.6	-12693.2	30	31	24.74592	6.54E-07	6.184375
g21410.t1	-636.264	-624.084	30	31	24.36092	7.99E-07	6.097601
g21437.t1	-9974.88	-9947.5	30	31	54.77059	1.35E-13	12.86823
g21489.t1	-1825.44	-1813.13	30	31	24.62461	6.97E-07	6.157038
g21541.t1	-3942.78	-3906.83	30	31	71.8951	0	Inf
g21616.t1	-3179.38	-3159.34	30	31	40.09368	2.42E-10	9.616058

Table A4.4C. Names and characteristics of genes marked by codeML as under positive selected in C. capreolus. *Gene codes and names are inferred using online blast tool of ncbi webpage. Potential outlier column indicates presence of 1 or more lineage specific amino acid mutations (LSAAM) with a BEB-score above 0.5 for class2a or class2b.*

gene	code	name	potential outlier	visual_check
g00068.t1	CCNB1	cyclin B1	TRUE	4 LSAAM with BEB>0.9 throughout gene
g00884.t1	PPP1R14B	phosphatase 1 regulatory inhibitor subunit 14B	TRUE	cluster of 4 adjacent LSAAM, of which 3 with BEB>0.8
g01006.t1	?	dehydrogenase family 3 member B1	FALSE	Many LSAAMs, but are surrounded by missing data in *C. capreolus*, suggestive of misalignment. (1 credible LSAAM)
g01347.t1	SCTR	secretin receptor	TRUE	cluster of 3 adjacent LSAAM with BEB>0.95
g01369.t1	IL16	pro-interleukin-16	FALSE	7 adjacent LSAAM, but surrounded by missing data in *C. capreolus*, suggestive of misalignment.
g01907.t1	ZFP91	zinc finger protein	FALSE	Many LSAAMs, but are surrounded by missing data in *C. capreolus*, suggestive of misalignment.
g01951.t1	TMCO1	transmembrane and coiled-coil domains 1	FALSE	Many LSAAMs, but are surrounded by missing data in *C. capreolus*, suggestive of misalignment.
g01978.t1	SRRM4	serine/arginine repetitive matrix 4	FALSE	Many LSAAMs, but are surrounded by missing data in *C. capreolus*, suggestive of misalignment.
g02746.t1	unknown	unknown	TRUE	3 adjacent LSAAM with BEB>0.5 for class2b
g03119.t1	PPEF2	protein phosphatase with EF-hand domain 2	TRUE	5 LSAAM, of which 3 adjacent, with BEB>0.5
g04152.t1	MYH8	heavy chain 8	TRUE	>22 clustered LSAAM (alongside many silent mutations; hence not marked by accelerated dN/dS tests). Clusters: 5 LSAAM in 12 codons, 5 LSAAM in 16 codons, one adjacent pair)
g04403.t1	KRT42	keratin type I cytoskeletal 42	TRUE	one cluster of 200 bp with many LSAAMs with BEB>0.5, one significant and one highly significant
g04906.t1	SLC12A5	solute carrier family 12 member 5	TRUE	section of 75 bp with 15 mutations, leading to 3 closely located LSAAM
g05061.t1	KMT2D	lysine methyltransferase 2D	FALSE	several LSAAMs, but none with BEB>0.5
g05187.t1	OPTN	optineurin	FALSE	Many LSAAMs, but are surrounded by missing data in all species except *C. capreolus*, suggestive of misalignment.
g05450.t1	BEAN1	brain expressed, associated with NEDD3	FALSE	Section with many LSAAMs, but are surrounded by missing data in *C. capreolus*, suggestive of misalignment.
g05700.t1	HDLBP	high density lipoprotein binding protein	FALSE	Misalignment
g06407.t1	SLC39A5	solute carrier family 39 member 5	TRUE	cluster of 3 adjacent LSAAM with BEB>0.5 (at the border of missing data, perhaps insertion in Hydropotes/*Capreolus* lineage)
g06580.t1	SYT2	synaptotagmin 2	TRUE	section of with >10 LSAAM in 30 codons, of which 8 with BEB>0.5, of which 3 significant
g06951.t1	SPATA21	spermatogenesis associated 21	FALSE	section with several LSAAMs, surrounded by missing data in *C. capreolus*, suggestive of misalignment
g07049.t1	OSGIN1	oxidative stress induced growth inhibitor 1	TRUE	6 LSAAM with BEB>0.5, of which 2 due to missing data, and the other 4 adjacent and significant

g07071.t1	GSE1	Gse1 coiled-coil protein	FALSE	3 adjacent LSAAM next to section of missing data in *C. capreolus*
g07241.t1	?	elongation of very long chain fatty acids protein 4	TRUE	cluster of 3 adjacent LSAAM, of which 2 with BEB>0.5
g07440.t1	EFCC1	EF-hand and coiled coil domain containing 1	FALSE	section with LSAAMs surrounded by missing data in *C. capreolus*
g07609.t1	ZNF444	zinc finger protein 444	FALSE	section with LSAAMs surrounded by missing data in *C. capreolus*
g07689.t1	AMPD2	adenosine monophoshate deaminase 2	FALSE	4 adjacent LSAAM next to section of missing data in *C. capreolus*
g07894.t1	?	eosinophil peroxidase	TRUE	5 LSAAM, of which a cluster of 3 adjacent, with BEB>0.5
g08423.t1	DDX42	DAED-box helicase 42	TRUE	cluster of 4 adjacent LSAAM with BEB>0.5, one significant
g08732.t1	?	complement C3	FALSE	section with LSAAMs surrounded by missing data in *C. capreolus*
g09357.t1	DMTF1	cyclin D binding myb like transcription factor 1	FALSE	section with LSAAMs surrounded by missing data in *C. capreolus*
g09440.t1	FAM189A2	family with sequence similarity 189 member A2	TRUE	7 LSAAM, of which a cluster of 4 adjacent, with BEB>0.5
g09476.t1	VPS13A	vacuolar protein sorting 13 homolog A	TRUE	7 LSAAM, of which 3 adjacent, with BEB>0.5
g10392.t1	TUBGCP2	tubulin gamma complex associated protein 2	FALSE	section with LSAAMs surrounded by missing data in *C. capreolus*
g11268.t1	?	serpin B3	TRUE	section of 50 bp with a cluster of 13 LSAAM with BEB>0.5, of which 10 adjacent
g11598.t1	ZNF783	zinc finger family member 783	TRUE	in first 200bp 9 LSAAM with BEB>0.5, of which a section of 50bp with a cluster of 7 LSAAM, of which 5 adjacent
g11691.t1	MDN1	AAA ATPase1	FALSE	19 LSAAM with BEB>0.5 in gene of 16kb, of which 8 surrounded by missing data in *C. capreolus*
g11849.t1	SPCS3	signal peptidase complex subunit 3	FALSE	cluster of 5 nearly adjacent LSAAM, of which 2 with BEB>0.5, most likely due to misalignment, because *C. pygargus* has identical sequence just upstream (with missing data for *C. capreolus*)
g12015.t1	SAMD5	sterile alpha motif domain containing 5	FALSE	section with LSAAMs surrounded by missing data in *C. capreolus*
g12222.t1	CABIN1	calcineurin binding protein 1	TRUE	3 LSAAM with BEB>0.5, of which 2 adjacent (in a cluster of 4 adjacent LSAAM)
g12484.t1	GNG4	G protein subunit gamma 4	FALSE	in first 100bp 8 LSAAM with BEB>0.5; including a cluster of 7 LSAAM out of 9 codons most likely due to misalignment, because identical sequence for *C. pygargus* upstream (missing data *C. capreolus*)
g12563.t1	PACS2	phosphofurin acidic cluster sorting protein 2	FALSE	LSAAMs in genomic region for which data is available only for *C. capreolus* and *C. pygargus*, so not characteristic for *C. capreolus*
g12719.t2	ADGRB2	adhesion G protein-coupled receptor B2	FALSE	section with LSAAMs surrounded by missing data in *C. capreolus*
g12882.t1	IRS2	insulin receptor substrate 2	FALSE	section with LSAAMs surrounded by missing data in *C. capreolus*
g13700.t1	MYOT	myotilin	FALSE	section with LSAAMs surrounded by missing data in *C. capreolus*
g13760.t1	TCF3	transcription factor 3	TRUE	cluster of 4 adjacent LSAAM with BEB>0.5
g13957.t1	TKT	transketolase	TRUE	4 LSAAM, of which a cluster of 3 adjacent and with BEB>0.95 (resulting from 9 adjacent nucleotide mutations)
g14096.t1	TTC9B	tetratricopeptide repeat domain 9B	TRUE	cluster of 3 adjacent LSAAM at end of sequence, of which 2 with BEB>0.95
g14583.t1	GNPTG	N-acetylglucosamine-1 phosphate transferase subunit gamma	TRUE	A cluster of 3 adjacent LSAAM with BEB>0.5, and a cluster of 4 adjacent LSAAM with BEB>0.5
g15865.t1	CCDC92	coiled coil domain containing 92	TRUE	A cluster of 3 adjacent LSAAM, of which 2 with BEB>0.5 and a stop codon
g16824.t1	ZAN	zonadhesin	TRUE	A cluster of 3 adjacent LSAAM, of which 2 with BEB>0.5
g17856.t1	RBBP7	RB binding protein 7, chromatin remodeling factor	FALSE	section with LSAAMs surrounded by missing data in *C. capreolus*

g18061.t1	unknown	unknown	FALSE	LSAAMs in genomic region for which data is available only for *C. capreolus* and *C. pygargus*, so not characteristic for *C. capreolus*
g18092.t1	unknown	unknown	TRUE	3 LSAAMs with BEB>0.5, of which a cluster of 2 adjacent
g18276.t1	peptidase 35	ubiquitin specific peptidase 35	FALSE	Section with many LSAAMs, but are surrounded by missing data in *C. capreolus*, suggestive of misalignment.
g18672.t1	TTC21B	tetratricopeptide repeat domain 21B	TRUE	7 LSAAMs with BEB>0.5, of which 3 adjacent
g18675.t1	SCN2A	sodium voltage-gated channel alpha subunit 2	TRUE	44 LSAAMs with BEB>0.5 (last section maybe unreliable)
g18676.t1	SCN3A	sodium voltage-gated channel alpha subunit 3	TRUE	40 LSAAMs with BEB>0.5 (last section maybe unreliable)
g18726.t1	NEB	nebulin	TRUE	30 LSAAMs with BEB>0.5
g18830.t2	COL11A2	collagen type XI alpha 2 chain	FALSE	misalignment, section with many mutations is in fact identical sequence occurs in *C. pygargus* where *C. capreolus* has missing data
g18911.t1	ESF1	ESF1 nucleolar pre-rRNA processing protein homolog	FALSE	missing data in *C. capreolus* leads to skewed estimate
g19145.t1	WAC	WW domain containing adaptor with coiled-coil	TRUE	4 adjacent LSAAM with BEB>0.5
g19891.t1	UTP20	small subunit processome component	TRUE	Definite candidate for positive selection: 8 non-adjacent LSAAM with BEB>0.5; one section of 70 bp with 5 LSAAM
g21058.t1	NMD3	NMD3 ribosome export adaptor	FALSE	high proportion of missing data in all species
g21142.t1	PRICKLE3	prickle planar cell polarity protein 3	FALSE	section with LSAAMs surrounded by missing data in *C. capreolus*
g21241.t1	PLXNA3	plexin A3	FALSE	paralog comparison
g21410.t1	SHROOM2	shroom family member 2	FALSE	high proportion of missing data in all species, might obscure codeml calculations
g21437.t1	DRP2	dystrophin related protein 2	FALSE	section with LSAAMs surrounded by missing data in *C. capreolus*
g21489.t1	DEK	DEK proto-oncogene	FALSE	section with LSAAMs surrounded by missing data in *C. capreolus*
g21541.t1	DCAF12L2	DDB1 and CUL4 associated factor 12-like protein 2	TRUE	1 LSAAM with BEB>0.5
g21616.t1	PLS3	plastin 3	TRUE	3 adjacent LSAAM with BEB>0.95

Table A4.5A. PAML's codeML likelihood ratio test (LRT) scores of genes marked by codeml as under positive selection in the species C. pygargus. *Ln0: log likelihood of neutral model (i.e. all codons: ω ≤ 1). Ln1: log likelihood of nested model with positive selection (i.e. some codons: ω > 1). Np0: number of parameters of neutral model. Np1: number of parameters of positive selection model. D-statistic: 2(ln1 – ln0). p-value: Chi-squared test p-value associated with D-statistics and 1 degree of freedom (np1 – np0). logp: negative log10 of p-value.*

gene	ln0	ln1	np0	np1	D	pvalue	logp
g00002.t1	-2075.65	-2061.94	30	31	27.40334	1.65E-07	6.782133
g01212.t1	-15847.3	-15821.8	30	31	50.86592	9.89E-13	12.00481
g06203.t1	-2622.19	-2610.37	30	31	23.65158	1.15E-06	5.937602
g06637.t1	-13444	-13399	30	31	89.88716	0	Inf
g06831.t1	-2070.1	-2054.18	30	31	31.85102	1.66E-08	7.778684
g06841.t1	-4929.93	-4910.72	30	31	38.41591	5.72E-10	9.242873
g09420.t1	-2224.84	-2211.85	30	31	25.97843	3.45E-07	6.461862
g10234.t1	-1618.95	-1601.32	30	31	35.2648	2.88E-09	8.54093
g20974.t1	-2470.57	-2449.15	30	31	42.85093	5.91E-11	10.2286
g21474.t1	-2437.78	-2405.53	30	31	64.49819	9.99E-16	15.00035

Table A4.5B. Names and characteristics of genes marked by codeML as under positive selection in the species C. pygargus. *Gene codes and names are inferred using online blast tool of ncbi webpage. Potential outlier column indicates presence of 1 or more lineage specific amino acid mutations (LSAAM) with a BEB-score above 0.5 for class2a or class2b.*

gene	name		logp	Potential outlier	visual_check
g00002.t1	APOL3	apolipoprotein L3	6.78	FALSE	No LSAAM present, erroneous LRT-score
g01212.t1	MAP1A	microtubule associated protein 1A	12.00	TRUE	section of 8 LSAAM (not clustered) with BEB>0.95 in close vicinity (but after stop codon)
g06203.t1	AK6	adenylate kinase 6	5.94	FALSE	all codons with BEB>0.5, of which 2 LSAAM with BEB>0.8. Real outlier is *C. capreolus*, which has a stopcodon.
g06637.t1	MUC2	mucin-2, oligomeric mucus gel-forming	Inf	TRUE	47 LSAAMs, of which 3 adjacent (next to an insertion in *C. pygargus*)
g06831.t1	RAB22A	RAS oncogene family	7.78	FALSE	paralog comparison
g06841.t1	PAXBP1	PAX3 and PAX7 binding protein	9.24	FALSE	many mutations in first 200 bp, but no data for this section for *C. capreolus*
g09420.t1	NAP1L1	assembly protein 1 like 1	6.46	TRUE	2 LSAAMs with BEB>0.5, of which one missing data in *C. capreolus*
g10234.t1	ZADH2	zinc binding alcohol dehydrogenase domain containing 2	8.54	TRUE	cluster of 5 LSAAM, of which 3 with BEB>0.5, possibly in insertion in *Capreolus*/Hydropotes lineage
g20974.t1	ZRSR2Y	CCCH-type zinc finger RNA-binding motif and serine/arginine rich 2Y-linked protein	10.23	FALSE	two LSAAM after an indel of one bp, but neither with BEB>0.5
g21474.t1	ZCCHC18	CCHC domain containing 18	15.00	FALSE	paralog comparison

Table A4.6A. Names and characteristics of genes with accelerated dN/dS rates. *Gene codes and names are inferred using online blast tool of ncbi webpage. Potential outlier column indicates presence of 1 or more lineage specific amino acid mutations (LSAAM).*

gene	code	name	Foreground branch	codeML-log p	potential outlier	Visual examination
g01940.t1	TMCC1	transmembrane and coiled-coil domain family 1	genus	0.56	TRUE	16 LSAAM
g16753.t1	DAGLB	diacylglycerol lipase beta	*C_pygargus*	NA	TRUE	one region with 6 LSAAM
g00678.t1	SF3B1	splicing factor 3b subunit 1	*C_capreolus*	0.49	TRUE	7 LSAAM in long gene with few changes
g01220.t1	MFAP1	microfibrillar associated protein 1	*C_capreolus*	1.01	TRUE	21 LSAAM, but including stopcodons
g05064.t1	TUBA1B	tubulin alpha 1b	*C_capreolus*	0.44	FALSE	paralog in C. elaphus
g05067.t1	TUBA1A	tubulin alpha 1a	*C_capreolus*	0.99	FALSE	35 LSAAM, but including stopcodon, and perhaps confounded by g05067.t1
g10584.t1	EEF2	translation elongation factor 2	*C_capreolus*	1.38	TRUE	13 LSAAM
g11717.t1	SLC16A7	solute carrier family 16 member 7	*C_capreolus*	0.88	TRUE	7 LSAAM
g18675.t1	SCN2A	sodium voltage-gated channel alpha subunit 2	*C_capreolus*	8.46	TRUE	around 30 LSAAM, plus very different last region
g18676.t1	SCN3A	sodium voltage-gated channel alpha subunit 3	*C_capreolus*	8.31	TRUE	21 LSAAM
g19134.t1	PCSK2	proprotein convertase subtilisin/kexin type 2	*C_capreolus*	3.13	TRUE	up to 30 LSAAM

Table A4.7. GO enrichment analysis. *InterPro terms identified as under selection by codeml branchsite tests with the species C. capreolus as foreground lineage.*

GO term	Function	Protein	p-value	adj. p-value
(IPR001696)	transmembrane transfer of sodium	SCN3A,SCN2A	1.26E-05	1.65E-02
(IPR019734)	protein-protein interactions	TTC21B,CABIN1,TTC9B	6.89E-05	4.27E-02
(IPR010526)	directed movement of sodium ions	SCN3A,SCN2A	1.52E-05	1.65E-02
(IPR003915)	interacting with Ca2+ ions	SCN3A,SCN2A	2.14E-05	1.86E-03
(IPR024583)	cytoplasmic domain in Na+ channel	SCN3A,SCN2A	1.01E-05	1.65E-03
(IRP000048)	non-covalent protein complex	MYH18,SCN3A,SCN2A	1.36E-05	1.65E-03

Table A4.8A.Non bootstrapped Ne estimates inferred by PSMC for C. capreolus

Time (ya)	Ne (10k)	# genomic regions
0	0.828763	5737.806
2270.945	0.828763	6003.02
4715.434	0.828763	6266.931
7346.265	0.828763	6527.242
10177.73	1.2264	4609.902
13225.3	1.2264	4835.953
16505.36	1.314904	4726.931
20035.76	1.314904	4948.556
23835.25	1.369732	4965.783
27924.98	1.369732	5182.538
32326.37	1.494386	4953.046
37063.83	1.494386	5159.514
42162.66	1.711326	4692.425
47650.26	1.711326	4886.34
53556.68	1.975925	4406.689
59913.78	1.975925	4588.479
66756.09	2.220665	4249.157
74120.21	2.220665	4420.115
82046.05	2.398132	4252.643
90576.75	2.398132	4412.928
99758.17	2.485088	4409.883
109640.3	2.485088	4556.512
120276.2	2.486963	4689.925
131723.5	2.486963	4814.816
144044.1	2.439017	5019.464
157304.9	2.439017	5109.161
171577.4	2.385686	5290.441
186938.8	2.385686	5329.805
203472.1	2.358425	5401.912
221266.7	2.358425	5380.617
240419.2	2.359571	5325.356
261032.8	2.359571	5239.92
283218.9	2.358256	5123.423
307097.6	2.358256	4969.847
332798.4	2.310301	4876.261
360459.9	2.310301	4641.153
390231.7	2.190853	4597.755
422275.3	2.190853	4258.233
456763	2.01464	4209.505
493882.2	2.01464	3749.048
533833.5	1.827529	3590.103
576832.9	1.827529	3033.606
623112.7	1.676017	2712.018
672923.1	1.676017	2150.082
726534	1.585467	1744.245
784235.2	1.585467	1291.079
846338.3	1.559582	941.1481
913179.9	1.559582	651.8857
985120.8	1.589412	430.2785
1062550	1.589412	280.6734
1145887	1.662047	170.4399
1235582	1.662047	105.3716
1332121	1.762187	59.85486
1436024	1.762187	35.16724
1547855	1.890978	18.77424
1668217	1.890978	10.51752
1797763	1.890978	5.606703
1937192	1.890978	2.833433

Table A4.8B.Non bootstrapped Ne estimates inferred by PSMC for C. pygargus

Time (ya)	Ne (10k)	# genomic regions
0	12.12489	5469.132
5873.233	12.12489	5789.408
12119.84	12.12489	6126.608
18763.85	12.12489	6481.386
25829.29	17.51649	4749.126
33343.88	17.51649	5030.524
41336.29	16.64186	5606.485
49836.99	16.64186	5933.966
58877.43	13.33314	7831.125
68492.71	13.33314	8271.038
78718.88	9.978776	11651.03
89594.76	9.978776	12259.56
101161.9	8.483758	15146.36
113463.8	8.483758	15879.46
126548.5	8.113093	17386.3
140464.1	8.113093	18179.48
155264.1	8.327995	18502.94
171004	8.327995	19313.94
187744.8	8.678485	19331.58
205550.2	8.678485	20147.33
224486	8.678106	20971.35
244625.9	8.678106	21798.19
266045.9	8.11289	24178.8
288826.3	8.11289	25009.31
313054.9	7.09226	29471.6
338823.2	7.09226	30229.89
366229.1	5.901994	37022.77
395377.2	5.901994	37473.66
426377.3	4.799074	46205.94
459346.9	4.799074	45854.81
494412.6	3.909845	55128.36
531706.4	3.909845	53235.94
571369.9	3.253614	60747.66
613554.6	3.253614	56639.77
658420.4	2.797239	60208.46
706137.3	2.797239	53833.16
756886.5	2.493578	52799.6
810861.1	2.493578	45044.02
868265.3	2.299978	40679.39
929318.3	2.299978	33008.27
994250.1	2.183493	27522.06
1063309	2.183493	21202.12
1136758	2.119937	16399.31
1214873	2.119937	11978.21
1297953	2.091629	8635.769
1386313	2.091629	5970.099
1480289	2.085603	4025.035
1580235	2.085603	2626.299
1686535	2.092368	1657.294
1799589	2.092368	1016.214
1919829	2.105159	599.0765
2047709	2.105159	343.1057
2183717	2.118791	188.0964
2328369	2.118791	99.80002
2482212	2.129558	50.49856
2645833	2.129558	24.56252
2819852	2.129558	11.36946
3004931	2.129558	4.992523

Table A4.9. Number of codeml PSGs vs genome quality and genetic diversity statistics.

Number of genes marked by codeml branch-site tests as putatively positively selected genes (PSGs) for various foreground branches, compared to genome wide heterozygosity (genome He) and genome assembly quality statistics (scaffold N50, contig N50, and average genome wide read depth).

Estimates of genome wide heterozygosity of red deer and white-tailed deer were generated in this study; the estimate for American bison and water buffalo are from Brüniche-Olsen et al. 2019 and Mintoo et al. 2019. Genome assembly quality statistics are obtained from NCBI or from the corresponding publications. Italic entries for authors and year refer to NCBI publication instead of journal publication.

Branch lengths refer to branch lengths in the exome tree (Fig. A.4.6A).

Foreground branch	Scaffold/ Contig N50	Branch Length	Average depth	Genome He (%)	# PSG	Authors	year
American bison	71926581997 1	NA	60	0.35	55	*Uni. of Maryland*	*2014*
Wisent	46900001453 0	NA	50	0.08	25	Wang et al.	2017
Bison genus					*15*		
Red deer	1073580067944	0.0032	74	0.14	99	Bana et al.	2018
Thorold deer	376937239627	0.0016	214	?	29	Chen et al.	2019
Wapiti	?6855	NA	40	?	39	Mizzi et al.	2017
Cervus genus					*10*		
Water buffalo	11721983522441509	NA	239	0.20	18	Low et al.	2019
Cape buffalo	240000043000	NA	90	0.06	26	Glanzmann et al.	2016
Bubalina subtribe					*12*		
European roe	104584167	0.0032	24	0.15	34	Kropatsch et al.	2013
Siberian roe	606722180310	0.0023	100	0.31	4	De Jong et al.	2020
Capreolus genus					*8*		
Mule deer	96789488	0.0040	26	?	86	*Canada Genome Enterprise*	*2018*
White-tailed deer	850721122019	0.0035	150	0.50	37	Seabury et al.	2011
Odocoileus genus					*12*		

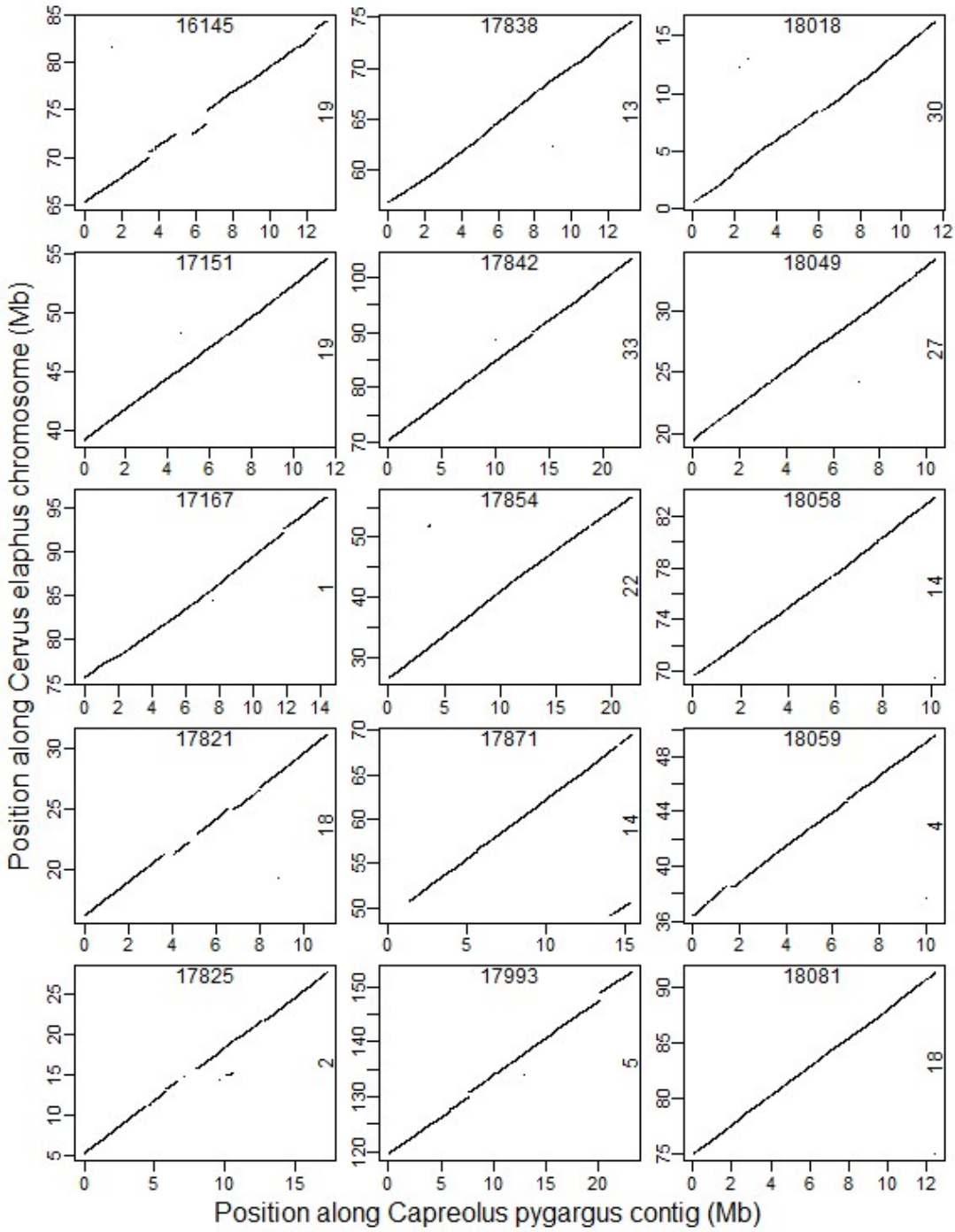


Fig. A4.1. Synteny analysis. *Dotplots showing output of whole genome alignments (using the software Lastz) of a random selection of C. pygargus contigs (minimum length: 10 Mb) against C. elaphus chromosomes. Before plotting, alignment results were filtered on sequence identity (>95%), alignment length (>300 bp), and number of hits per subject_ID (>500). Numbers at the top of the panels denote C. pygargus contigs; numbers at the right hand side denote C. elaphus chromosomes. The alignments indicate conserved synteny.*

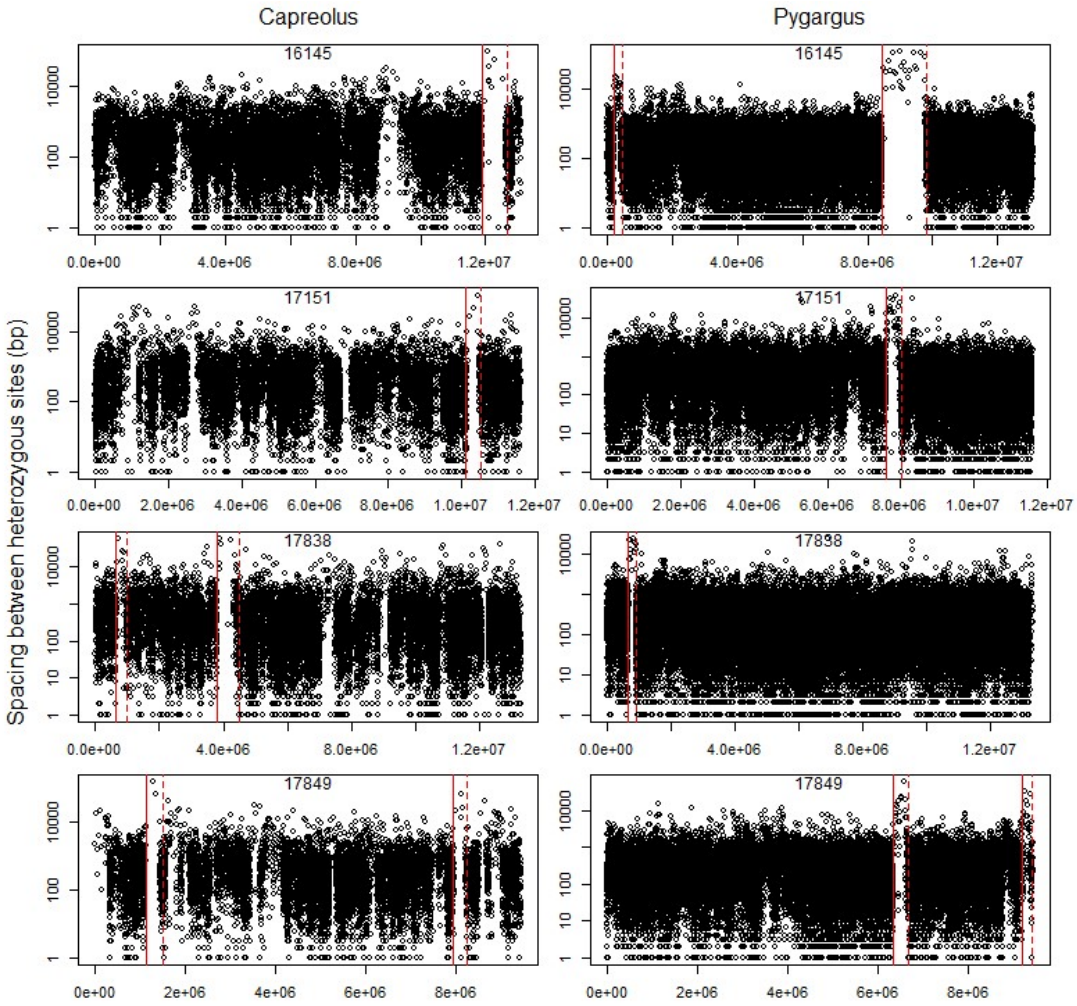


Fig. A4.2A*. Spacing between heterozygous sites in C. capreolus genome (left) and C. pygargus genome (right). Solid red line: start of low genetic diversity region. Dashed red line: end of low genetic diversity region. Shown are contigs which contain regions with low genetic diversity. No overlap of low genetic diversity regions is observed between both sister species.*

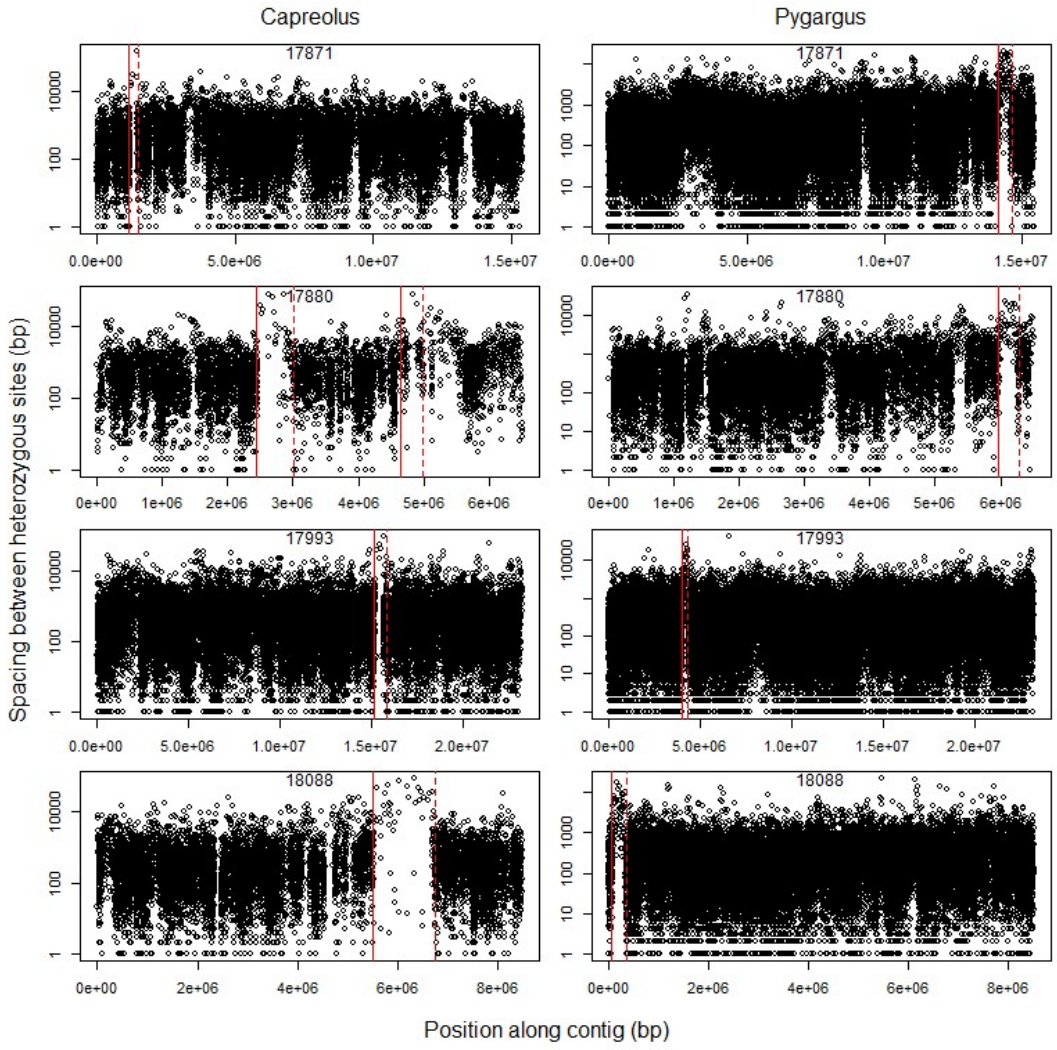


***Fig. A4.2A cont**. Spacing between heterozygous sites in C. capreolus genome (left) and C. pygargus genome (right). Solid red line: start of low genetic diversity region. Dashed red line: end of low genetic diversity region. Shown are contigs which contain regions with low genetic diversity. No overlap of low genetic diversity regions is observed between both sister species.*

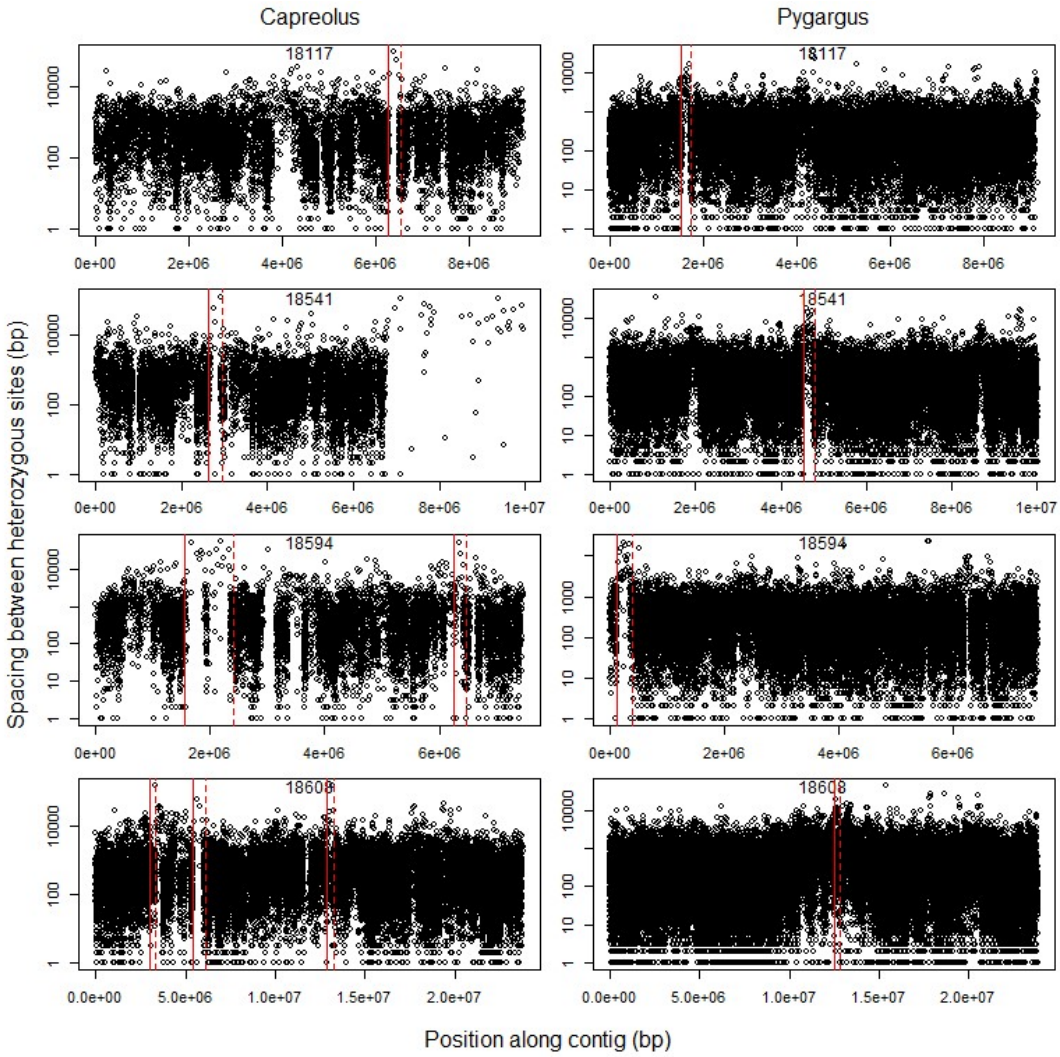


Fig. A4.2A cont. *Spacing between heterozygous sites in C. capreolus genome (left) and C. pygargus genome (right). Solid red line: start of low genetic diversity region. Dashed red line: end of low genetic diversity region. Shown are contigs which contain regions with low genetic diversity. No overlap of low genetic diversity regions is observed between both sister species.*

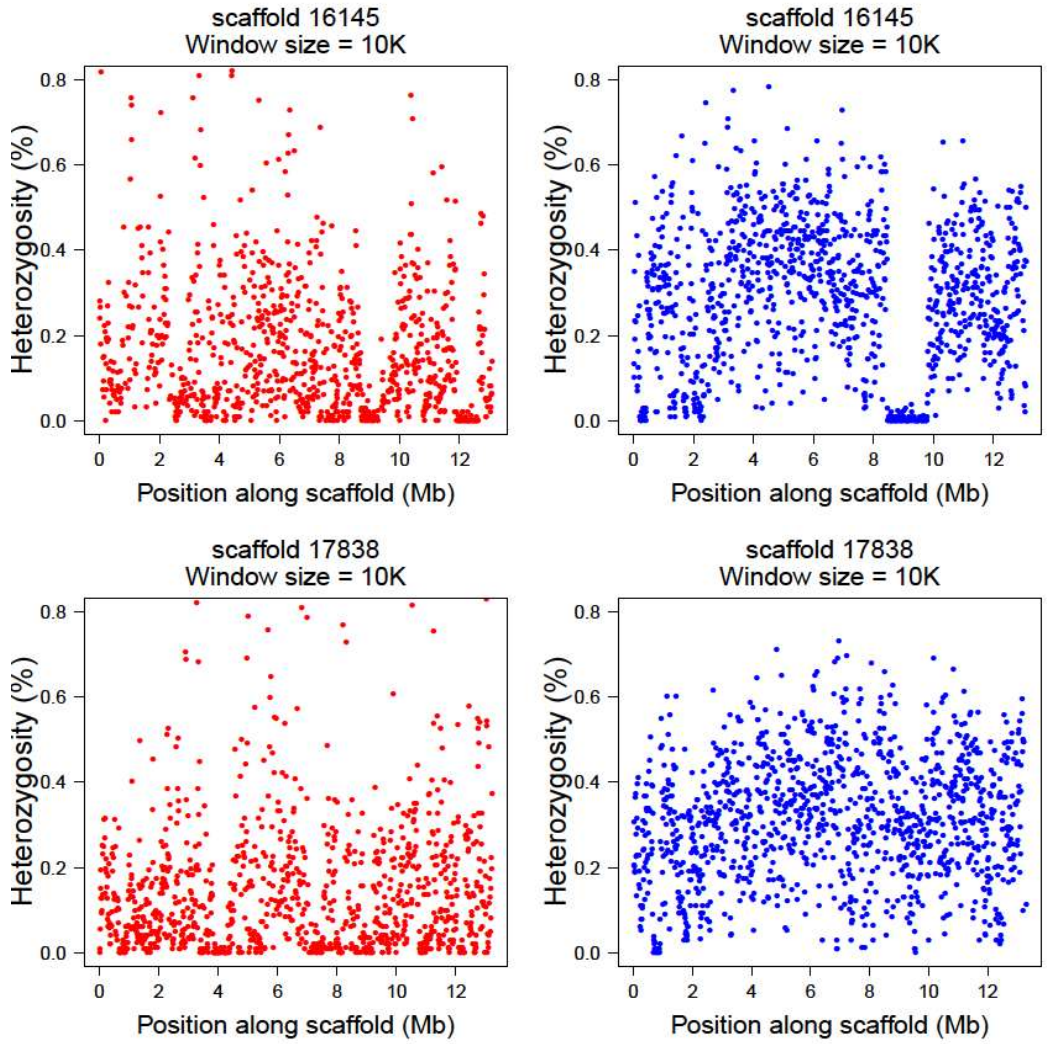


Fig. A4.2B. Sliding window heterozygosity estimates for scaffolds 16145 and 17838. *Sliding window heterozygosity analyses of C. capreolus (left, red) and C. pygargus (right, blue) genome confirm that regions marked by He-spacing analyses (see Fig. A4.2A) contain few heterozygous sites. Shown are here two scaffolds which according to the He-spacing analyses contain low diversity regions (see Fig. A4.2A).*

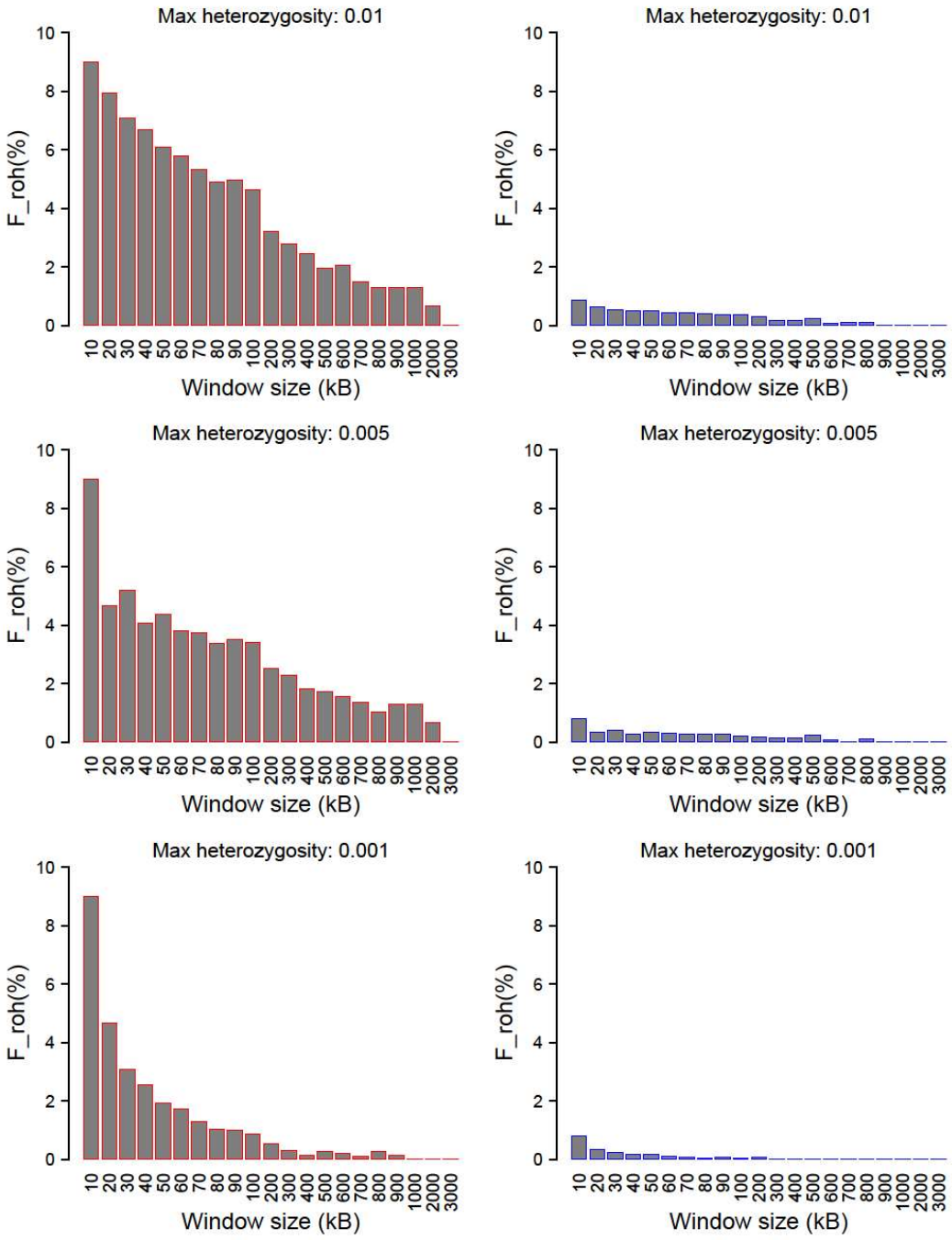


Fig. A4.2C. F_roh. *Percentage of genome with stretches of low heterozygosity (<0.01%, <0.005% and <0.001%) – i.e. runs of homozygosity (ROH) – within the C. capreolus genome (left, red) and the C. pygargus genome (right, blue), given various sizes of non-overlapping windows. E.g.: a F_roh score of 0.1% for a window size of 100Kb and a max heterozygosity of 0.01%, indicates that 0.1% of non-overlapping windows of 100Kb length have a heterozygosity equal or below 0.01%. Excluded from the analysis are windows with more than 20% missing data, and windows from contigs 19446 and 547919, which contain relatively low heterozygosity levels and are therefore possibly non-autosomal. Note that contrary to plink ROH analyses, adjacent ROH windows have not been combined into longer windows.*

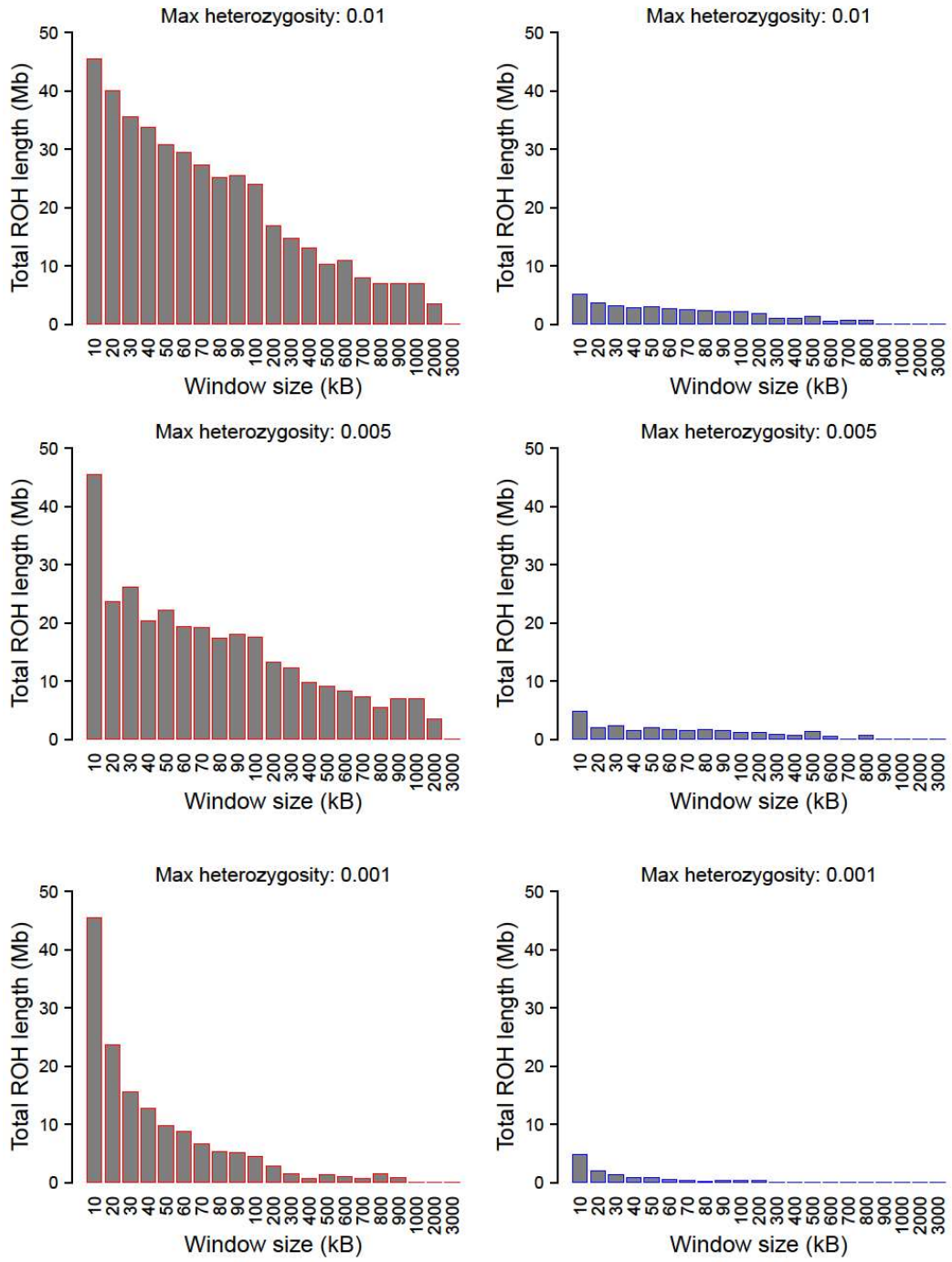


Fig. A4.2D. Total ROH length. *Idem as Fig. A4.2C, but showing combined length of runs of homozygosity (ROH) rather than proportion of the genome. In some case higher window size can unintuitively lead to slighlty higher total ROH lengths, because total ROH length is summed over window sized. For example, if a 650 kb stretch of low heterozygosity causes the average window heterozygosity to be below the threshold for both an overlapping 700 Kb and an overlapping 800 Kb window, the length will be recorded as respectively 700 Kb and 800 Kb.*

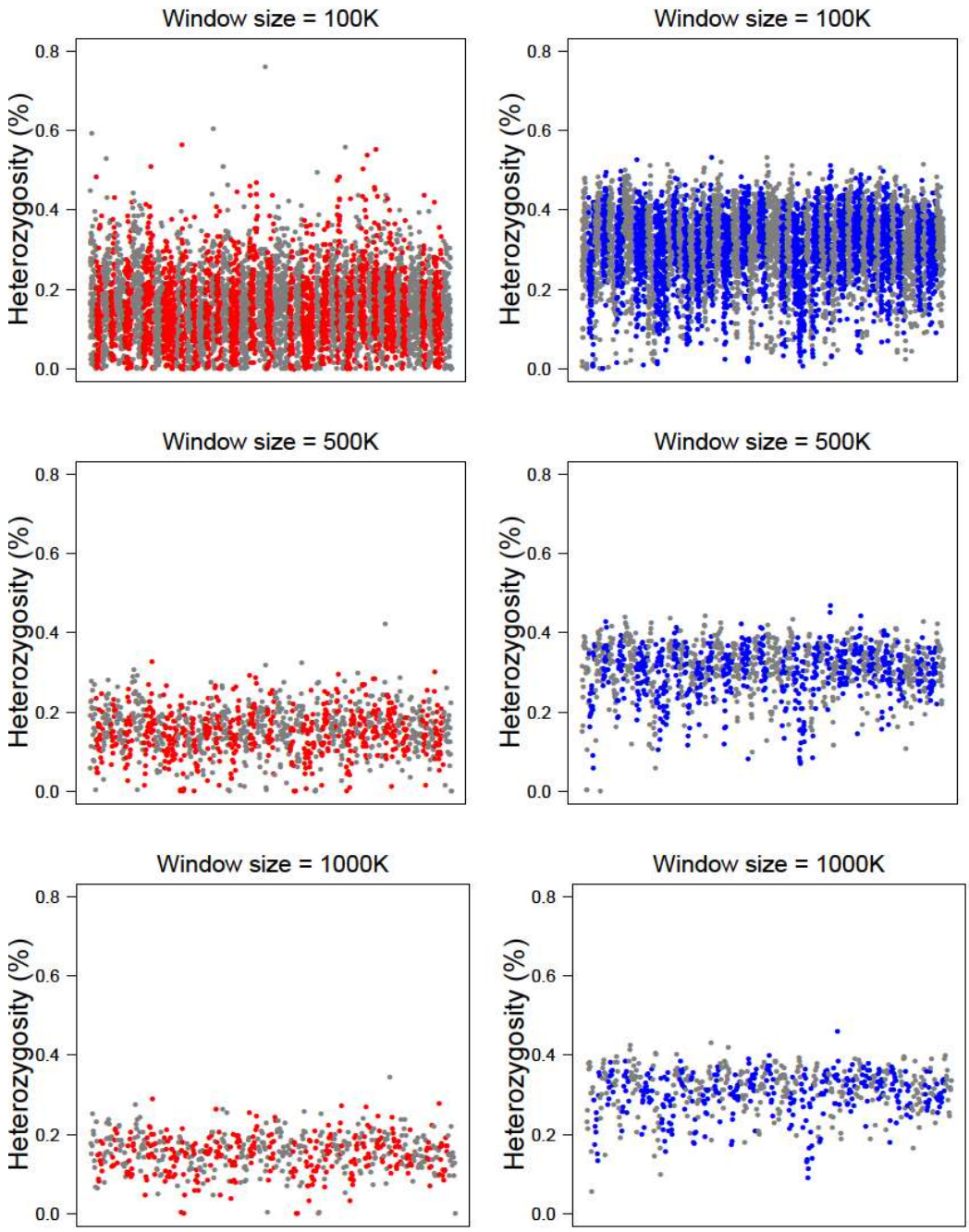


Fig. A4.2E. Sliding window heterozygosity. *Heterozygosity estimates (He) of non-overlapping windows of various sizes (0.1Mb, 0.5Mb and 1Mb) for C. capreolus (left, red) and C. pygargus (right, blue). As expected, the within-species variation of heterozygosity among windows depends on the size of the window size. Due to the lower genome wide heterozygosity of C. capreolus, more windows have heterozygosity levels close to zero, explaining (partly) the difference in Froh score between the two Capreolus species (Fig. A4.2B-C).*

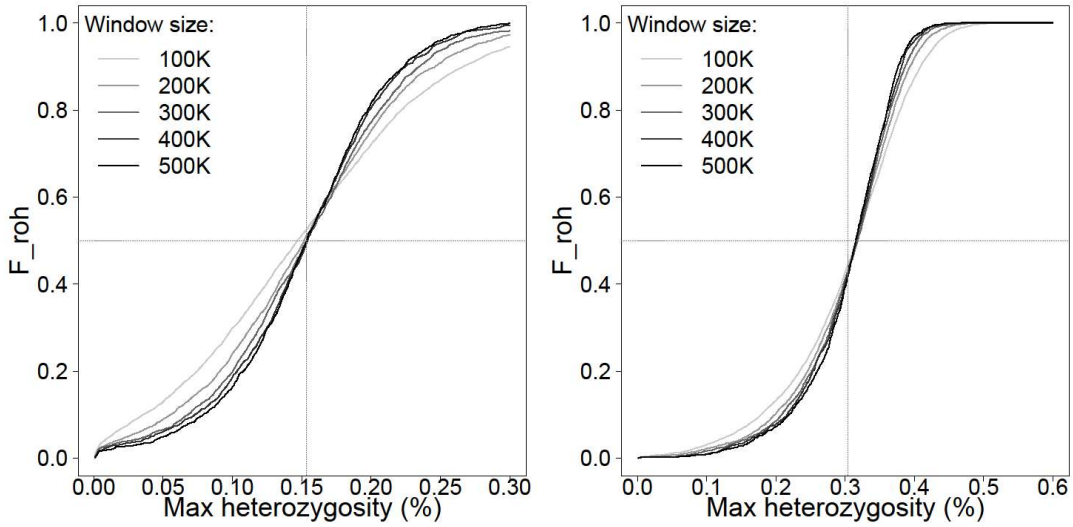


Fig. A4.2F. F_roh versus maximum heterozygosity threshold. *Plots showing the dependency of the F_roh estimate on the maximum heterozygosity threshold setting, given non-overlapping windows of various sizes, for C. capreolus (left) and C. pygargus (right, blue). As expected, F_roh equals around 0.5 if the maximum heterozygosity threshold (i.e. maximum level of heterozygosity used to define a window as a run of homozygosity) is set to the mean genome wide estimate.*

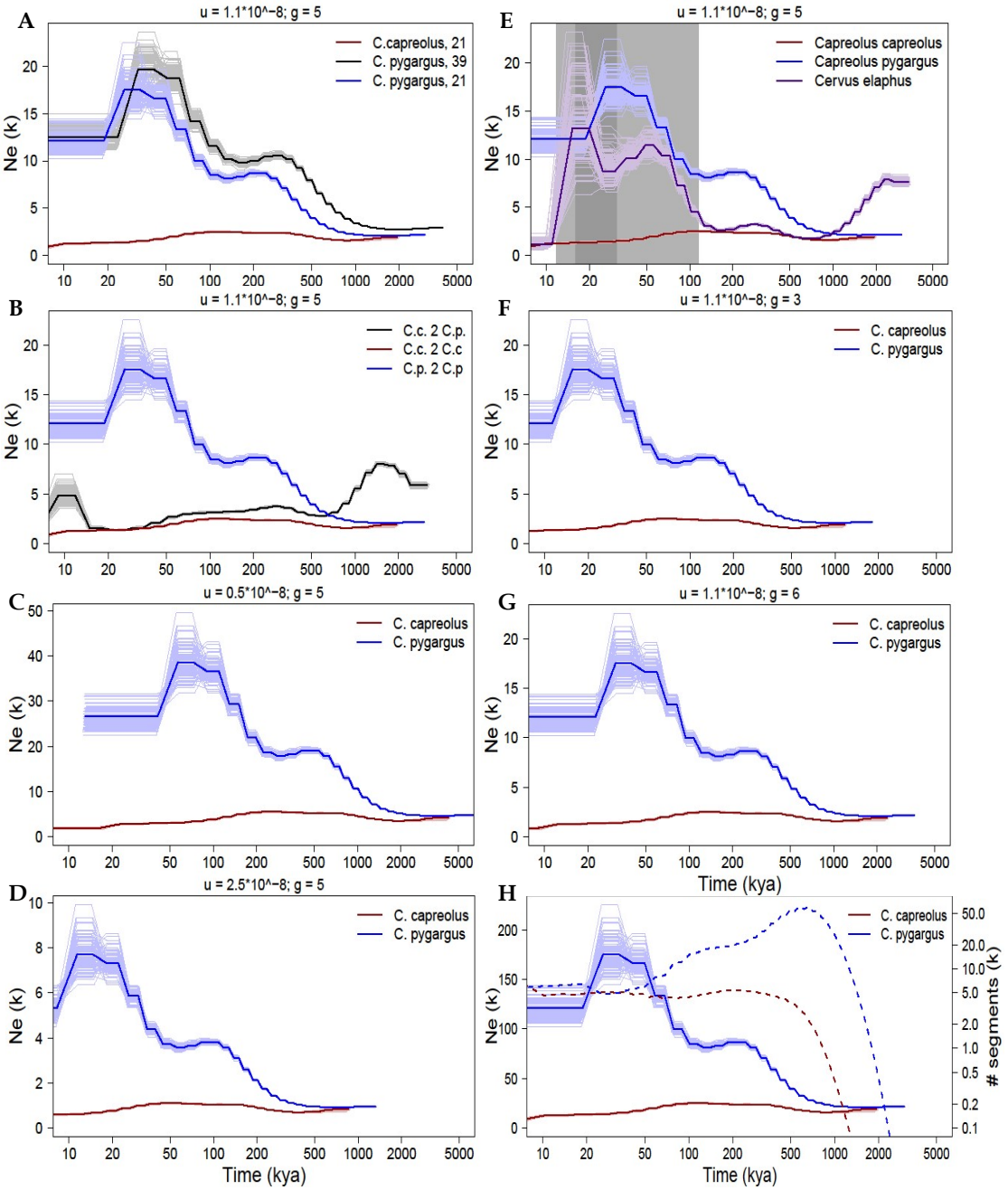


Figure A4.3.PSMC output under various settings. *Scale in y-axes is 10k rather than 1k, as falsely indicated by the y-axes labels in panels A-G. (**A**). The effect of downsampling the C. pygargus dataset from on average read depth of 39 (mean read depth of C. pygargus dataset) to an average read depth of 21 (mean read depth of C capreolus dataset) on historic Ne estimates of C. pygargus. (**B**). Effect of crossmapping C. capreolus reads on historic Ne estimates of C. capreolus. (**C-D**). Effect of generation specific mutation rates on historic Ne estimates of C. pygargus and C. capreolus. Note the different scales on the y-axes. (**E**). Demographic histories of C. capreolus compared to Cervus elaphus. Generation time of C. elaphus is set to 7 years (Coulson et al, 1998, Microsatellites reveal heterosis in red deer). (**F-G**). The effect of generation time on historic Ne estimates of C. pygargus and C. capreolus. (**H**). Number of genomic regions (dashed lines) with inferred TMRCA estimates.*

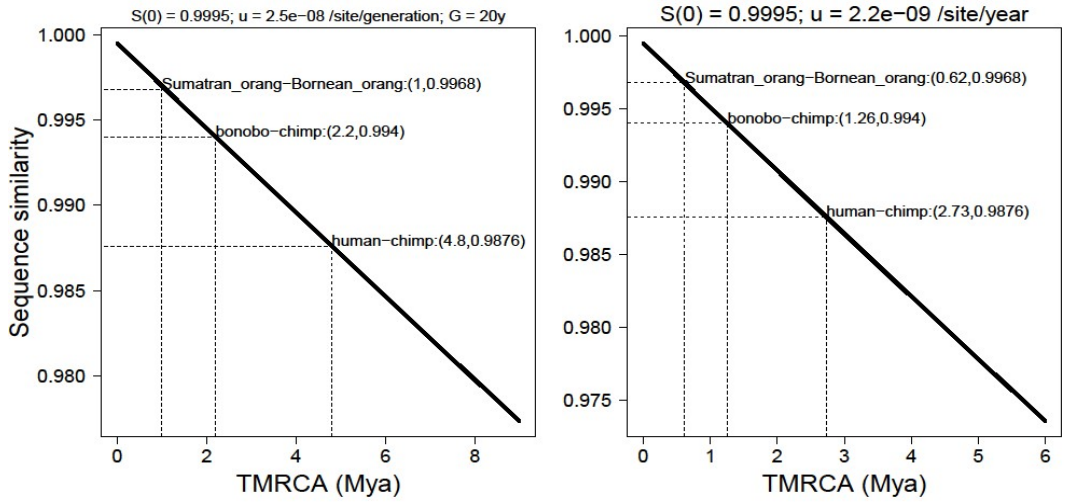


Figure A4.4.TMRCA estimates outputted by the random walk Markov chain model for great ape species pairs. *Dependency of TMRCA estimates of great ape species on mutation rate. Left: mutation rate of $2.5 \cdot 10^{-8}$ mutations per site per generation, with a generation time of 20 years. Right: mutation rate of $0.22 \cdot 10^{-8}$ mutations per site per year. The different outcomes of both approaches are caused by the non-linear relationship between mutation rate per year and mutation rate per generation.*

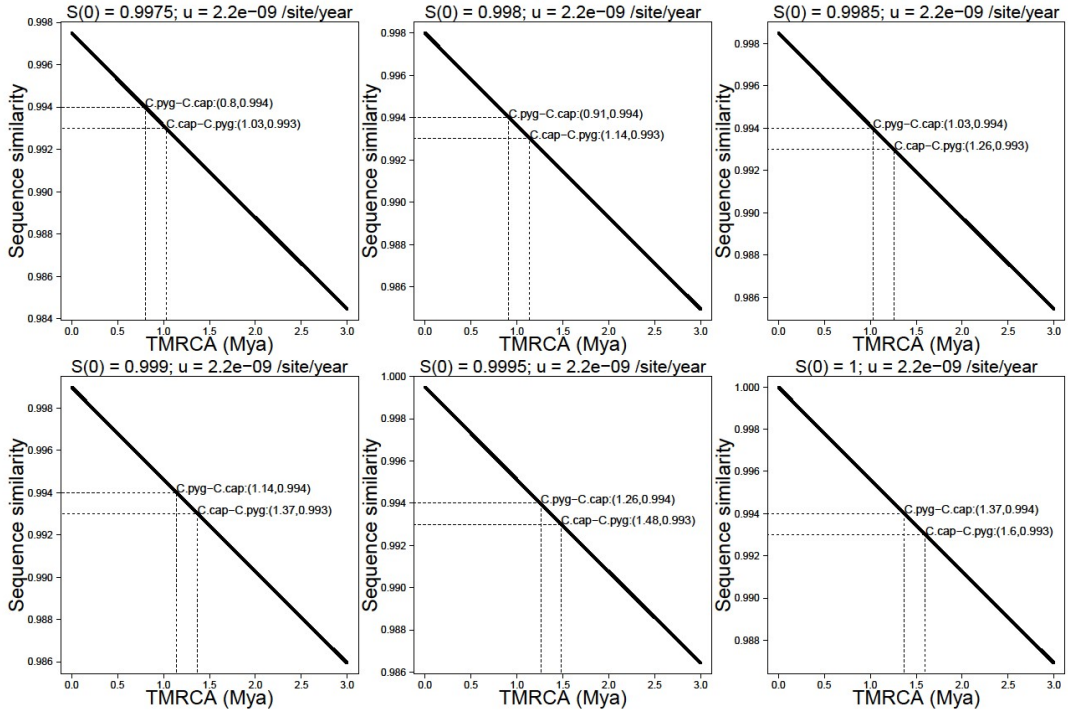


Figure A4.5.TMRCA estimates outputted by the random walk Markov chain model under various settings. *Dependency of TMRCA estimates of C. pygargus and C. capreolus on start (S_0) and end (S_n =0.993 or S_n = 0.994) sequence similarity estimates. The mutation rate is set to $2.2 \cdot 10^{-9}$ mutations per site per year. The start sequence similarity defines the sequence similarity of both sister population after fixation/ loss of standing variation.*

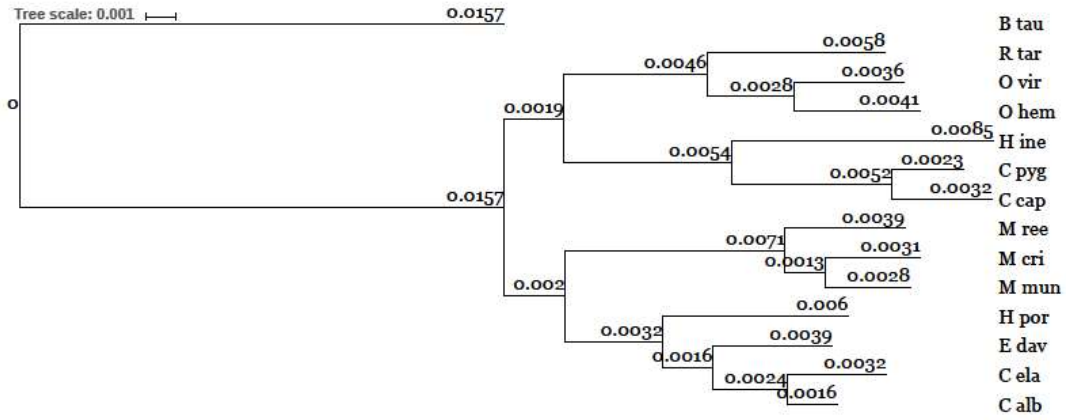


Figure A4.6A. *Maximum likelihood species tree of cervids based on full exomes with 100 bootstrap replications using the software RaxML (Stamatakis, 2014) with Bos taurus (cattle) as outgroup, and with partitioning into first and second codon positions vs third codon positions. All nodes have 100% bootstrap support. Labels indicate branch lengths. I generated this phylogenetic tree to verify the gene alignments. Tree plot generated on iTOL webpage (Letunic & Bork, 2019, Interactive Tree of Life (iTOL) v4: recent updates and new developments).*

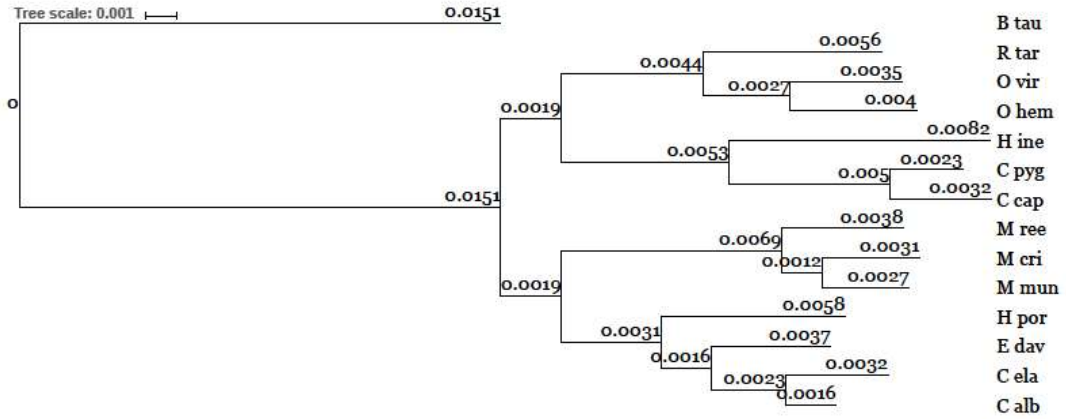


Figure A4.6B. *Maximum likelihood species tree of cervids based on 3rd codon positions with 100 bootstrap replications using the software RaxML (Stamatakis, 2014) with Bos taurus (cattle) as outgroup. All nodes have 100% bootstrap support. Labels indicate branch lengths. I generated this phylogenetic tree to verify the gene alignments. Tree plot generated on iTOL webpage (Letunic & Bork, 2019, Interactive Tree of Life (iTOL) v4: recent updates and new developments).*

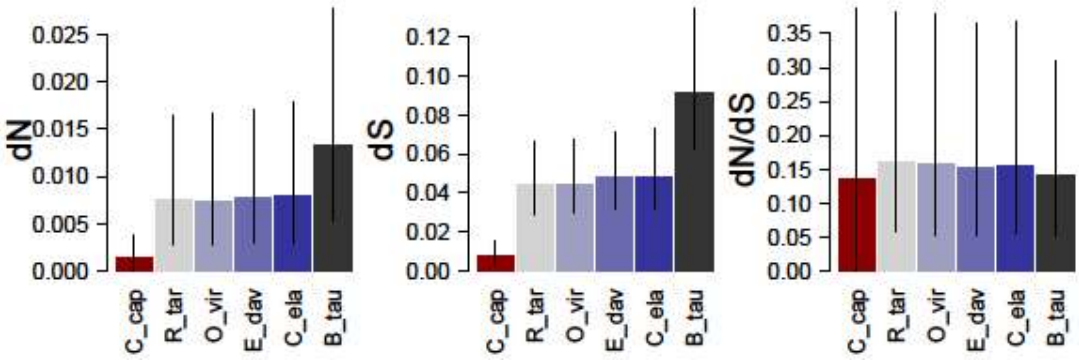


Fig A4.7. Pairwise median dN/dS scores*. Barplots showing distribution of dN and dS values, calculated using PAML's yn00, for pairwise comparisons between C. pygargus and 5 other cervid species as well as cattle, based on a dataset of up to 14,512 genes. Bar heights indicate median gene specific dN, dS and dN/dS values. Error bars indicate 25% and 75% percentiles.C_cap = C. capreolus (western roe deer), R_tar = Rangifer = tarandus (reindeer), O_vir = Odocoileus virginianus (white tailed deer), E_dav = Elaphurus davidianus (Pere David's deer), C_ela = Cervus elaphus (red deer), B_tau = Bos taurus (cattle).*

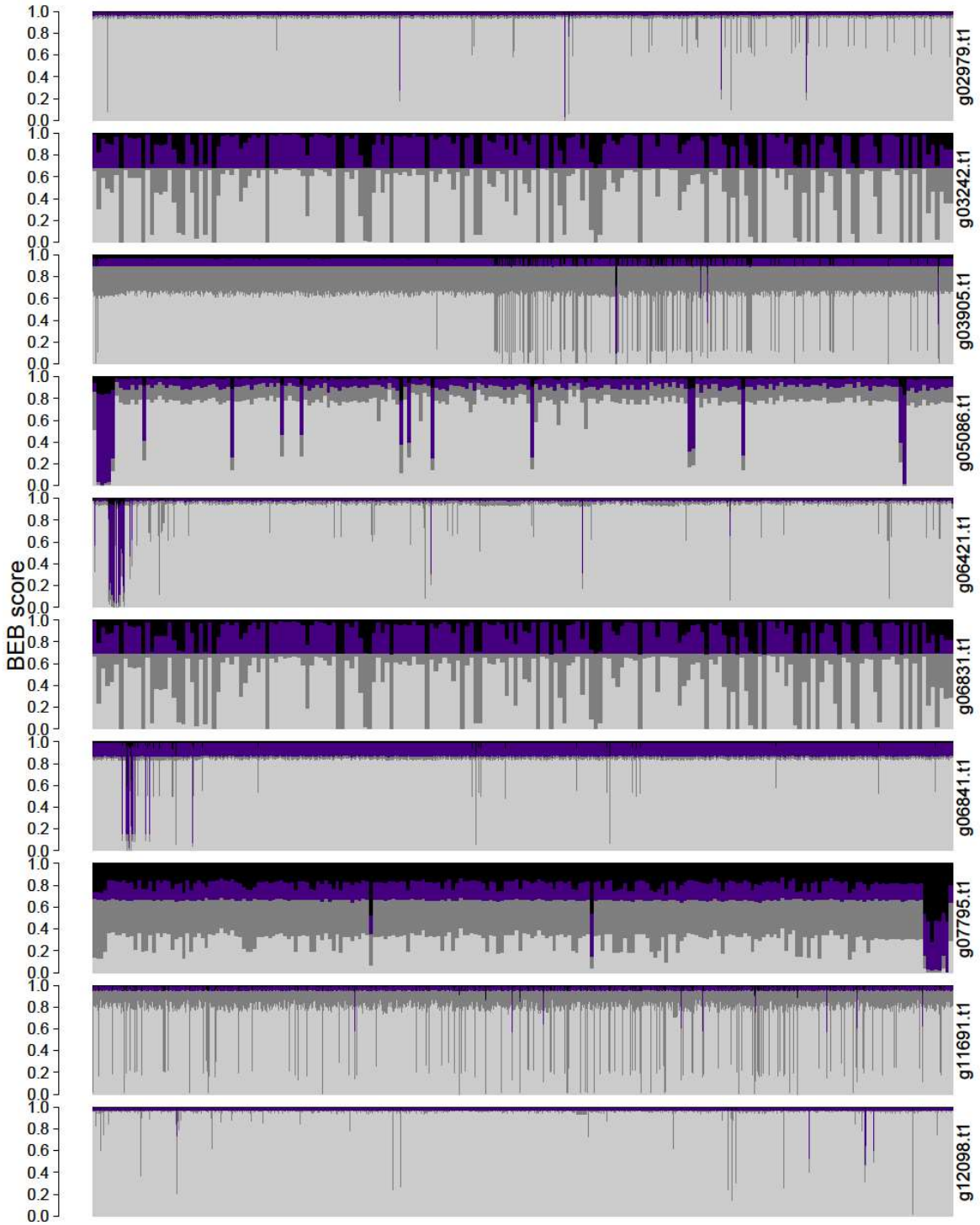


Fig A4.8. Bayesian probability of site class per codon, outputted by codeML branch site test, for genes outputted by codeML as having experienced episodic positive selection in the genus Capreolus*. x-axis: position along gene, y-axis:BEB-score. Lightgrey: $\omega_{for} < 1$ and $\omega_{back} < 1$. Darkgrey: $\omega_{for} = 1$ and $\omega_{back} = 1$. Purple: $\omega_{for} > 1$ and $\omega_{back} < 1$. Black: $\omega_{for} > 1$ and $\omega_{back} = 1$. Note: for long genes (>2000 bp) colours of individual codons might be lost.*

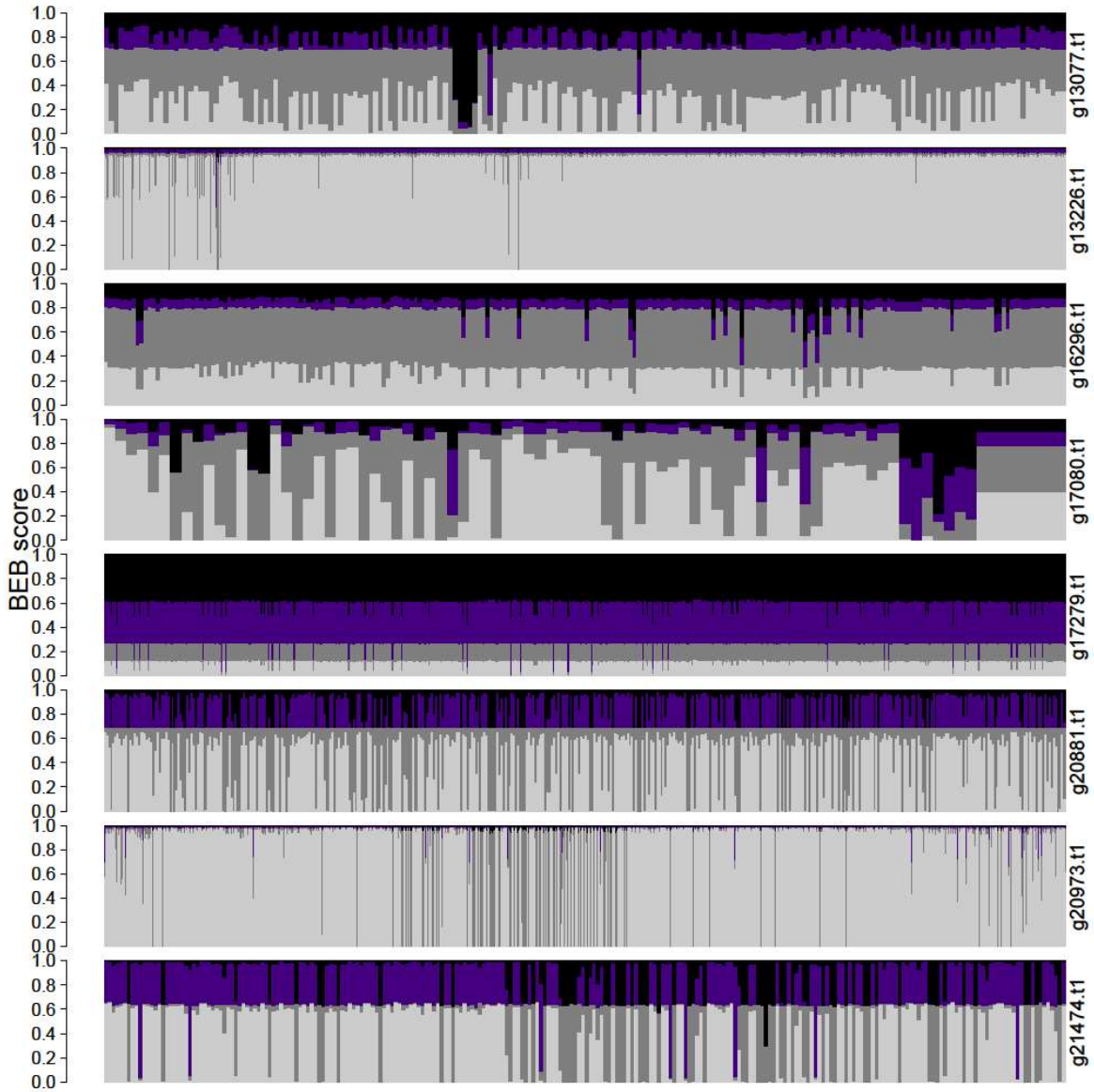


Fig A4.8 cont. Bayesian probability of site class per codon, outputted by codeML branch site test, for genes outputted by codeML as having experienced episodic positive selection in the genus Capreolus*. x-axis: position along gene, y-axis:BEB-score. Lightgrey: $\omega_{for} < 1$ and $\omega_{back} < 1$. Darkgrey: $\omega_{for} = 1$ and $\omega_{back} = 1$. Purple: $\omega_{for} > 1$ and $\omega_{back} < 1$. Black: $\omega_{for} > 1$ and $\omega_{back} = 1$. Note: for long genes (>2000 bp) colours of individual codons might be lost.*

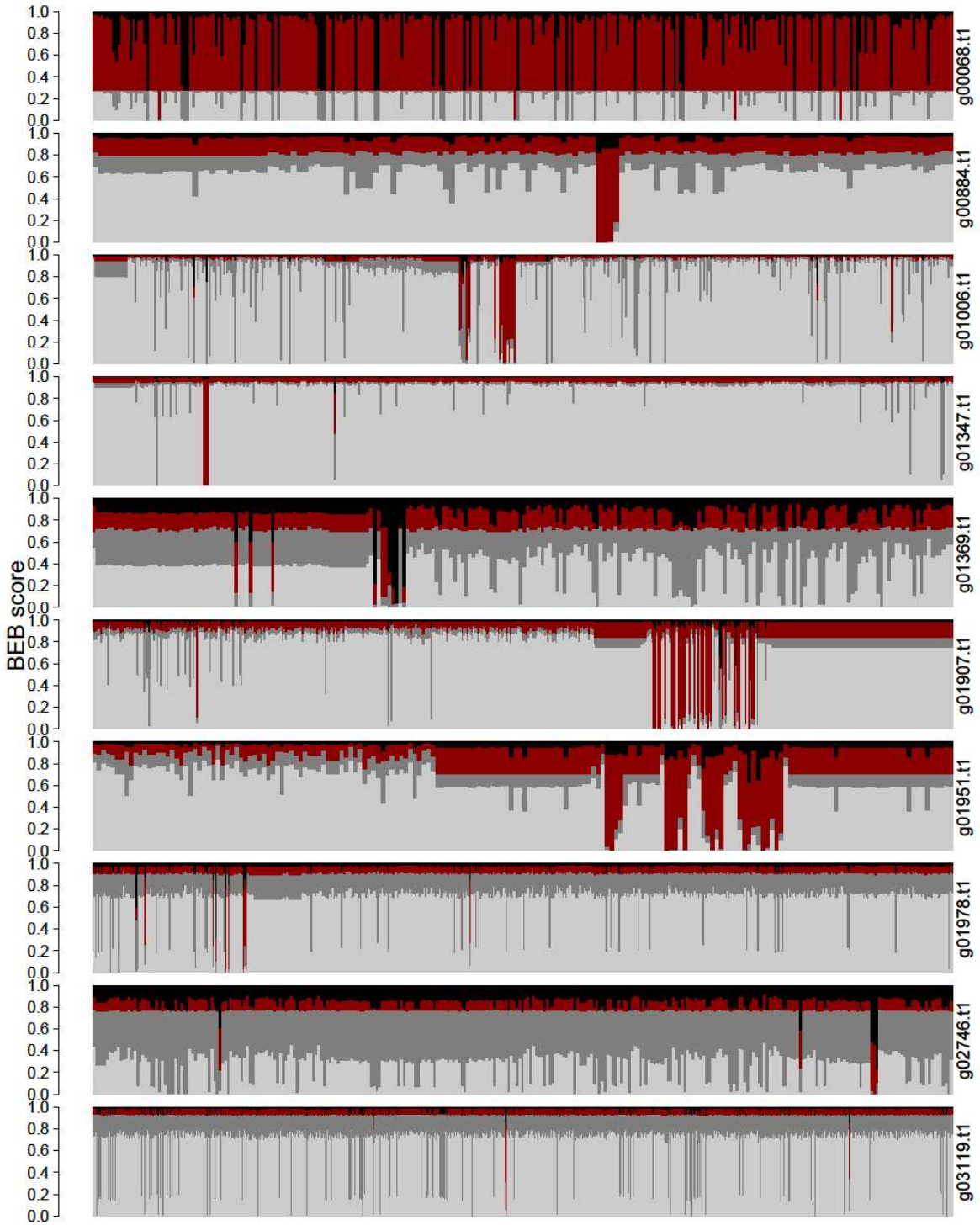


Fig A4.9. Bayesian probability of site class per codon, outputted by codeML branch site test, for genes outputted by codeML as having experienced episodic positive selection in the species C. capreolus*. x-axis: position along gene, y-axis:BEB-score. Lightgrey: $\omega_{for} < 1$ and $\omega_{back} < 1$. Darkgrey: $\omega_{for} = 1$ and $\omega_{back} = 1$. Red: $\omega_{for} > 1$ and $\omega_{back} < 1$. Black: $\omega_{for} > 1$ and $\omega_{back} = 1$. Note: for long genes (>2000 bp) colours of individual codons might be lost.*

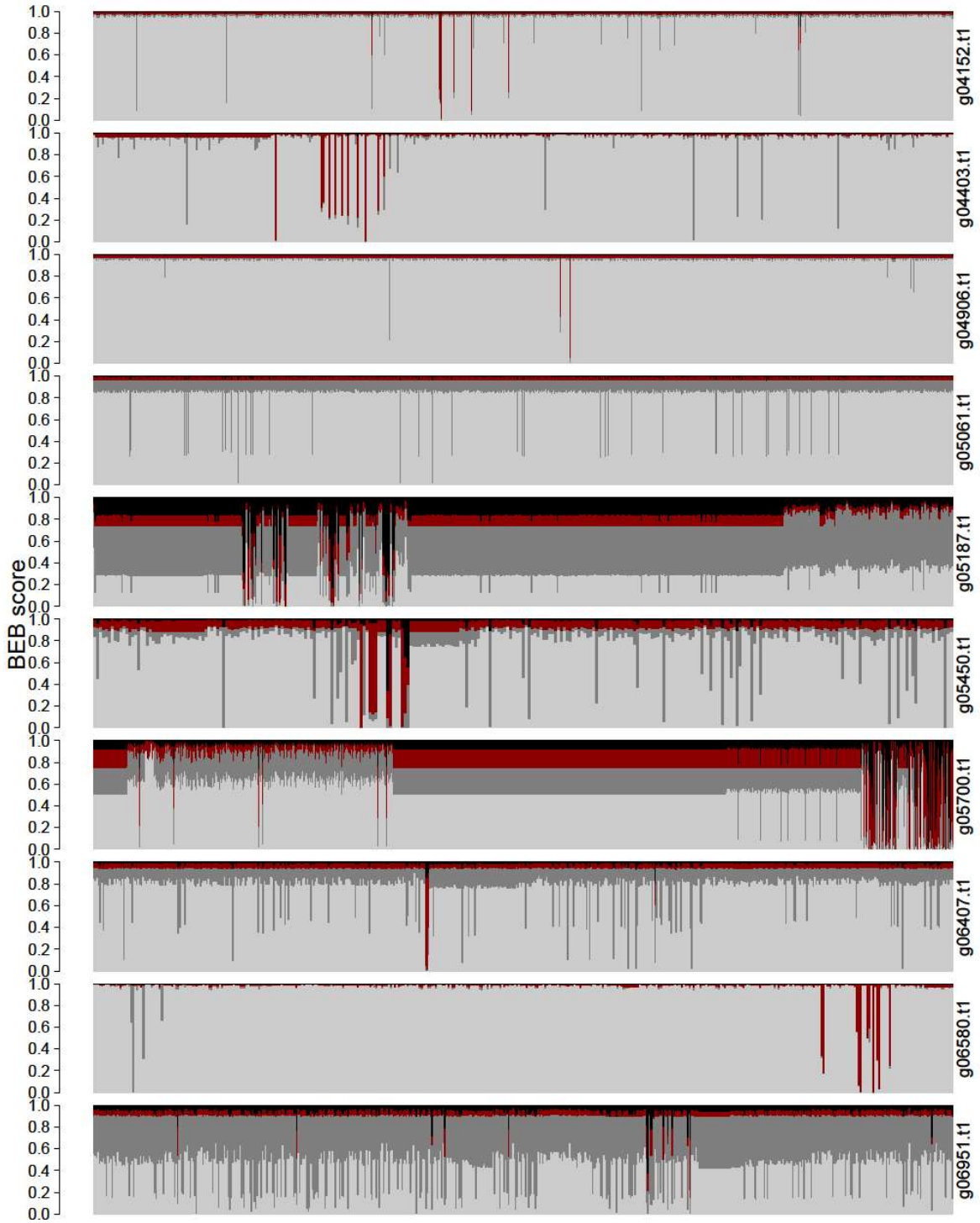


Fig A4.9 cont. Bayesian probability of site class per codon, outputted by codeML branch site test, for genes outputted by codeML as having experienced episodic positive selection in the species C. capreolus*. x-axis: position along gene, y-axis:BEB-score. Lightgrey: $\omega_{for} < 1$ and $\omega_{back} < 1$. Darkgrey: $\omega_{for} = 1$ and $\omega_{back} = 1$. Red: $\omega_{for} > 1$ and $\omega_{back} < 1$. Black: $\omega_{for} > 1$ and $\omega_{back} = 1$. Note: for long genes (>2000 bp) colours of individual codons might be lost.*

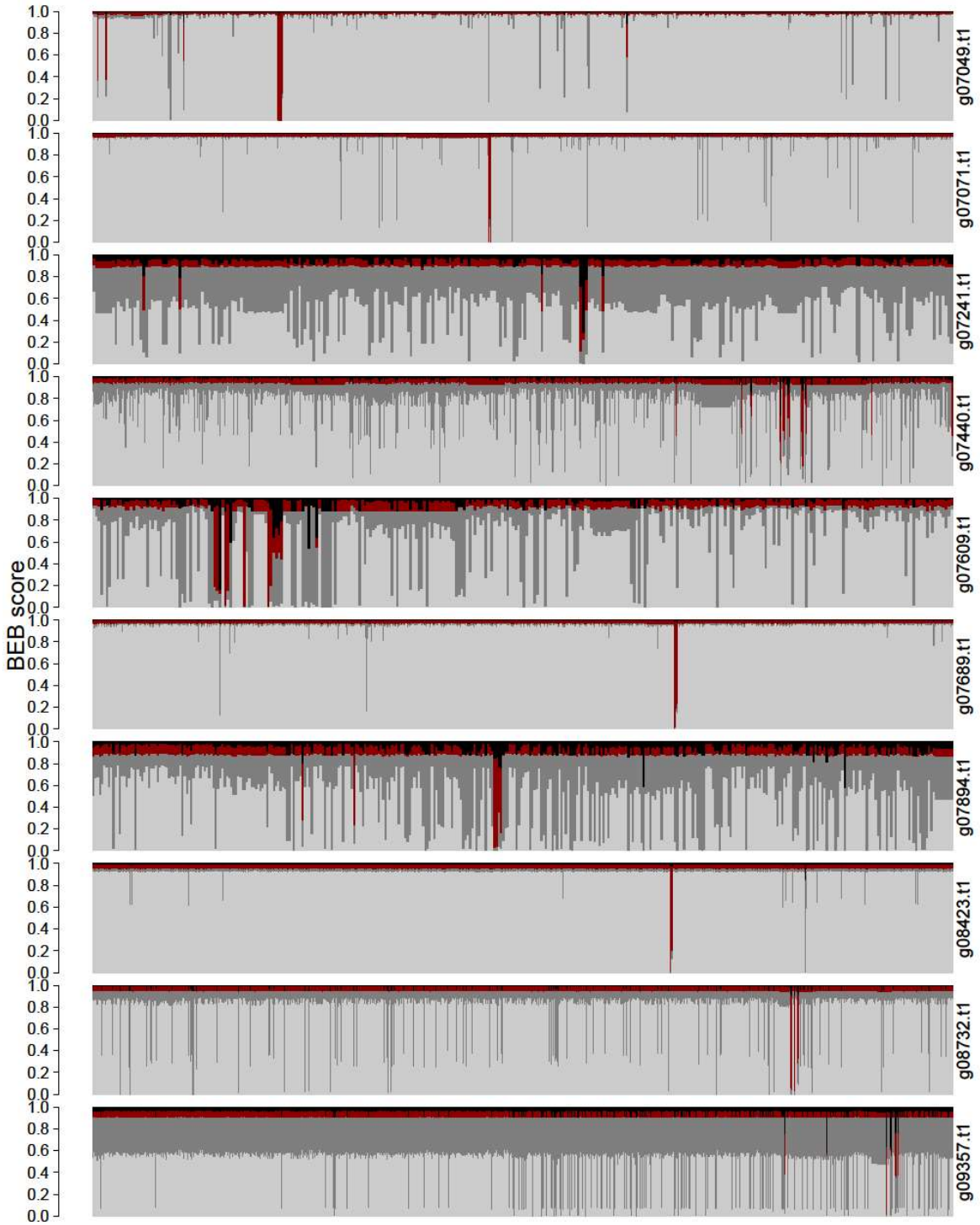


Fig A4.9 cont. Bayesian probability of site class per codon, outputted by codeML branch site test, for genes outputted by codeML as having experienced episodic positive selection in the species C. capreolus*. x-axis: position along gene, y-axis:BEB-score. Lightgrey: $\omega_{for} < 1$ and $\omega_{back} < 1$. Darkgrey: $\omega_{for} = 1$ and $\omega_{back} = 1$. Red: $\omega_{for} > 1$ and $\omega_{back} < 1$. Black: $\omega_{for} > 1$ and $\omega_{back} = 1$. Note: for long genes (>2000 bp) colours of individual codons might be lost.*

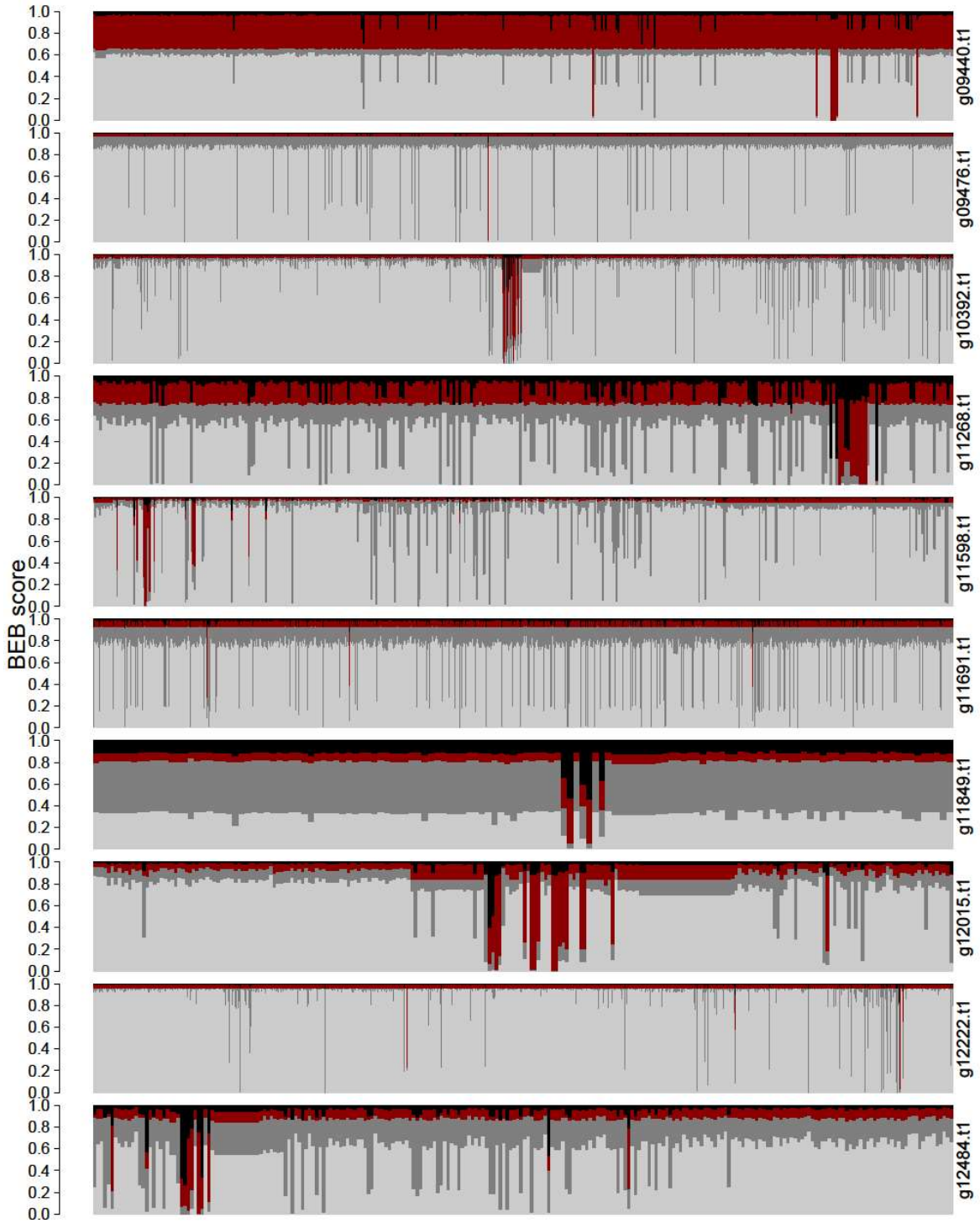


Fig A4.9 cont. Bayesian probability of site class per codon, outputted by codeML branch site test, for genes outputted by codeML as having experienced episodic positive selection in the species C. capreolus*. x-axis: position along gene, y-axis:BEB-score. Lightgrey: $\omega_{for} < 1$ and $\omega_{back} < 1$. Darkgrey: $\omega_{for} = 1$ and $\omega_{back} = 1$. Red: $\omega_{for} > 1$ and $\omega_{back} < 1$. Black: $\omega_{for} > 1$ and $\omega_{back} = 1$. Note: for long genes (>2000 bp) colours of individual codons might be lost.*

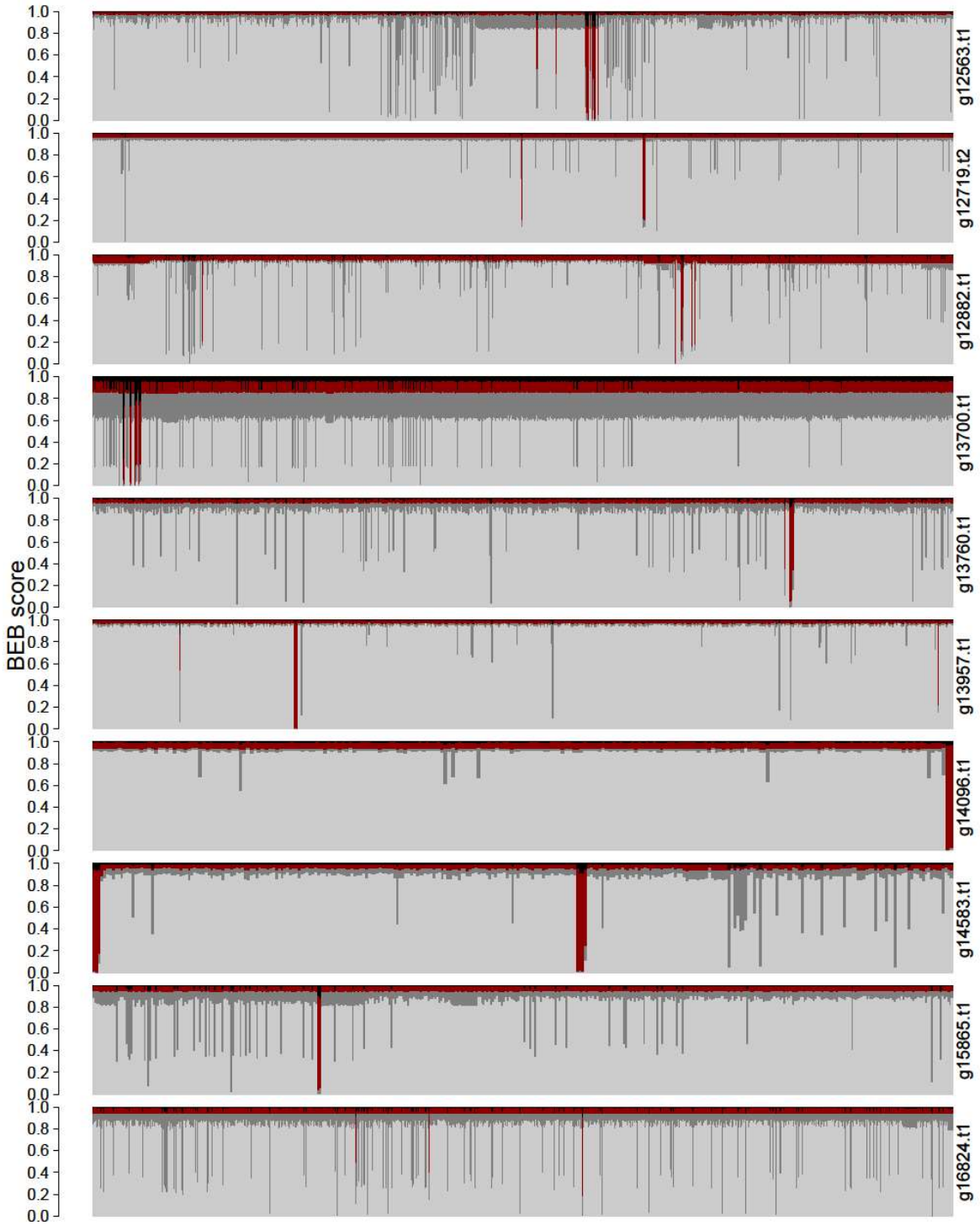


Fig A4.9 cont. Bayesian probability of site class per codon, outputted by codeML branch site test, for genes outputted by codeML as having experienced episodic positive selection in the species C. capreolus*. x-axis: position along gene, y-axis:BEB-score. Lightgrey: $\omega_{for} < 1$ and $\omega_{back} < 1$. Darkgrey: $\omega_{for} = 1$ and $\omega_{back} = 1$. Red: $\omega_{for} > 1$ and $\omega_{back} < 1$. Black: $\omega_{for} > 1$ and $\omega_{back} = 1$. Note: for long genes (>2000 bp) colours of individual codons might be lost.*

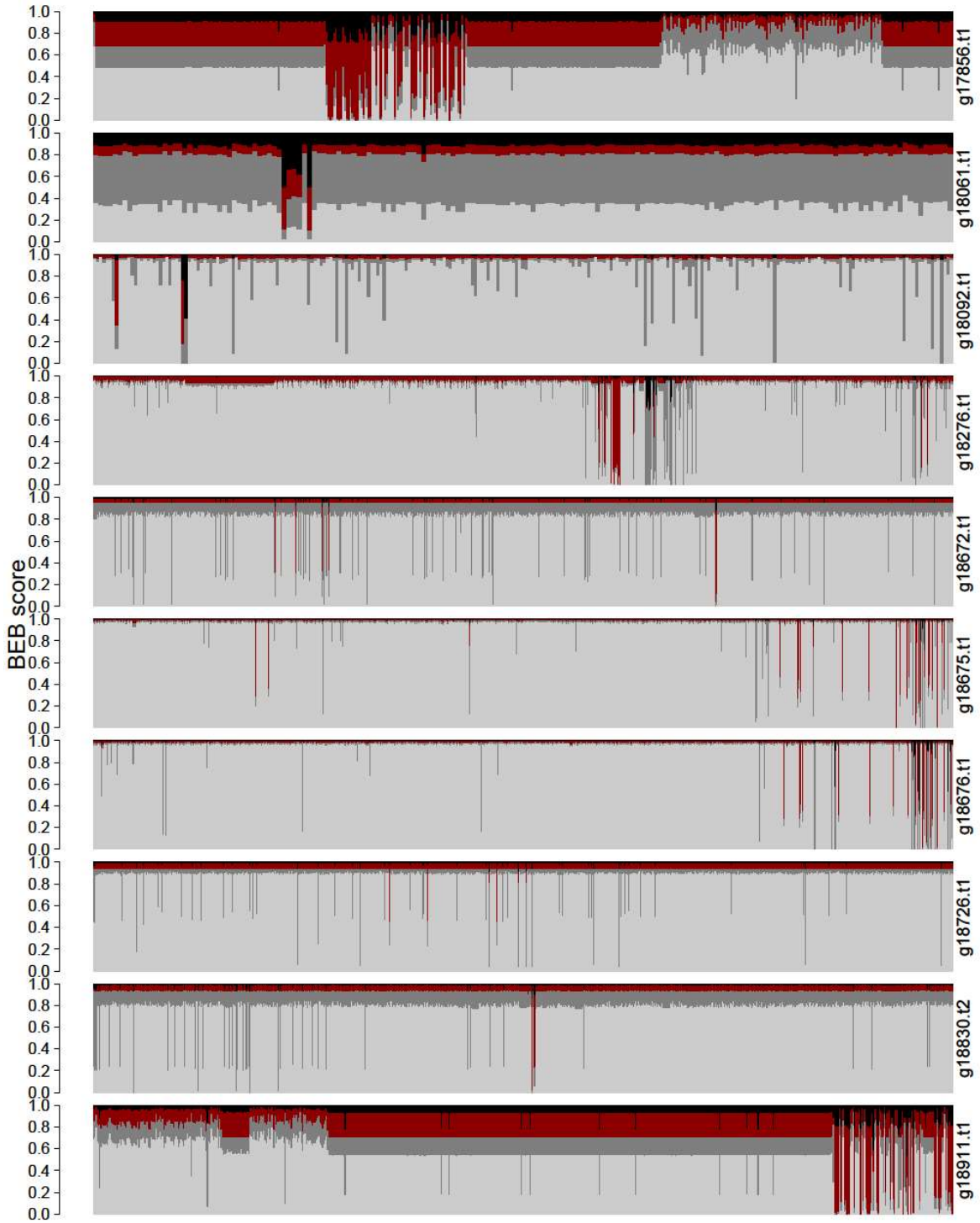


Fig A4.9 cont. Bayesian probability of site class per codon, outputted by codeML branch site test, for genes outputted by codeML as having experienced episodic positive selection in the species C. capreolus*. x-axis: position along gene, y-axis:BEB-score. Lightgrey: $\omega_{for} < 1$ and $\omega_{back} < 1$. Darkgrey: $\omega_{for} = 1$ and $\omega_{back} = 1$. Red: $\omega_{for} > 1$ and $\omega_{back} < 1$. Black: $\omega_{for} > 1$ and $\omega_{back} = 1$. Note: for long genes (>2000 bp) colours of individual codons might be lost.*

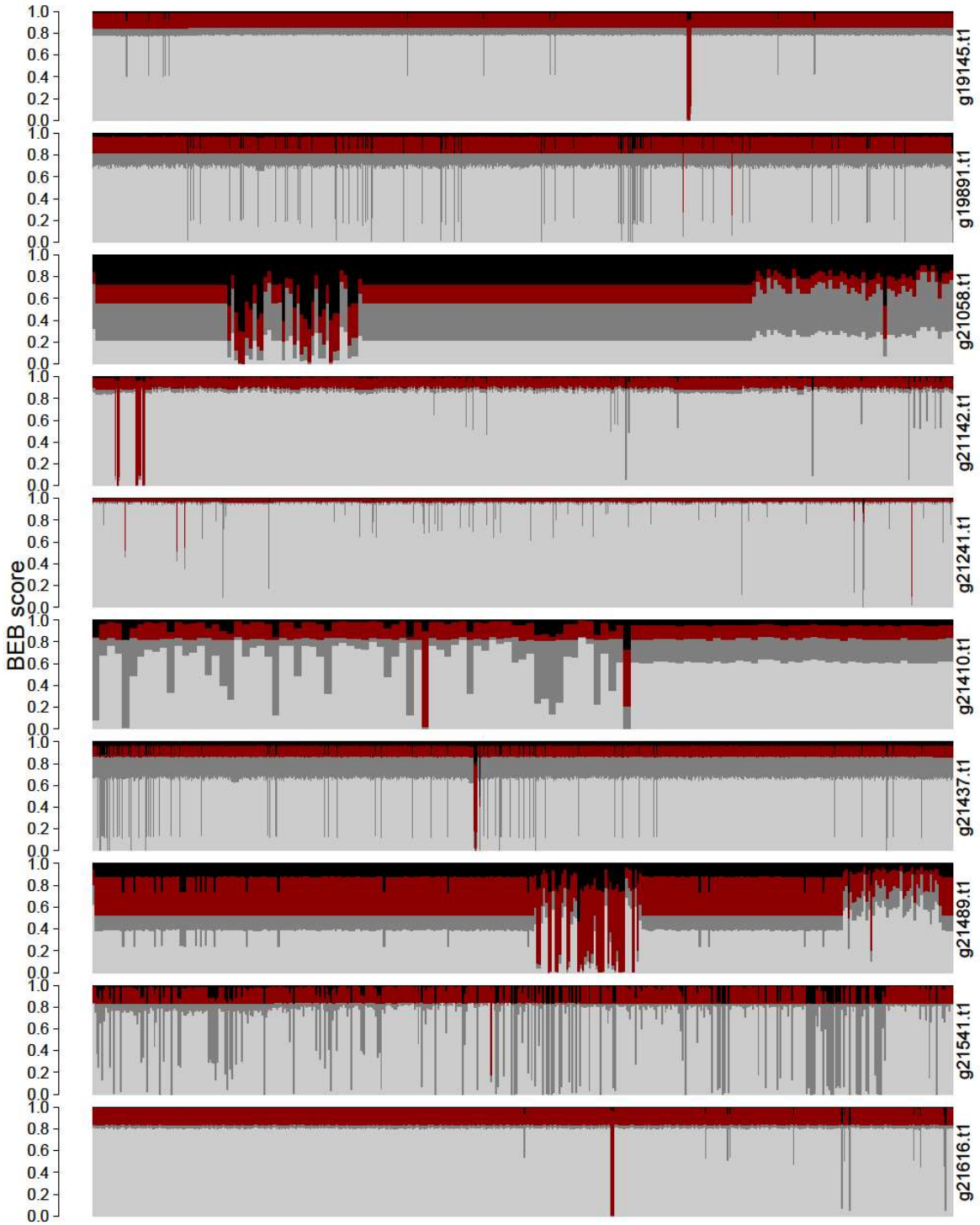


Fig A4.9 cont. Bayesian probability of site class per codon, outputted by codeML branch site test, for genes outputted by codeML as having experienced episodic positive selection in the species C. capreolus*. x-axis: position along gene, y-axis:BEB-score. Lightgrey: $\omega_{for} < 1$ and $\omega_{back} < 1$. Darkgrey: $\omega_{for} = 1$ and $\omega_{back} = 1$. Red: $\omega_{for} > 1$ and $\omega_{back} < 1$. Black: $\omega_{for} > 1$ and $\omega_{back} = 1$. Note: for long genes (>2000 bp) colours of individual codons might be lost.*

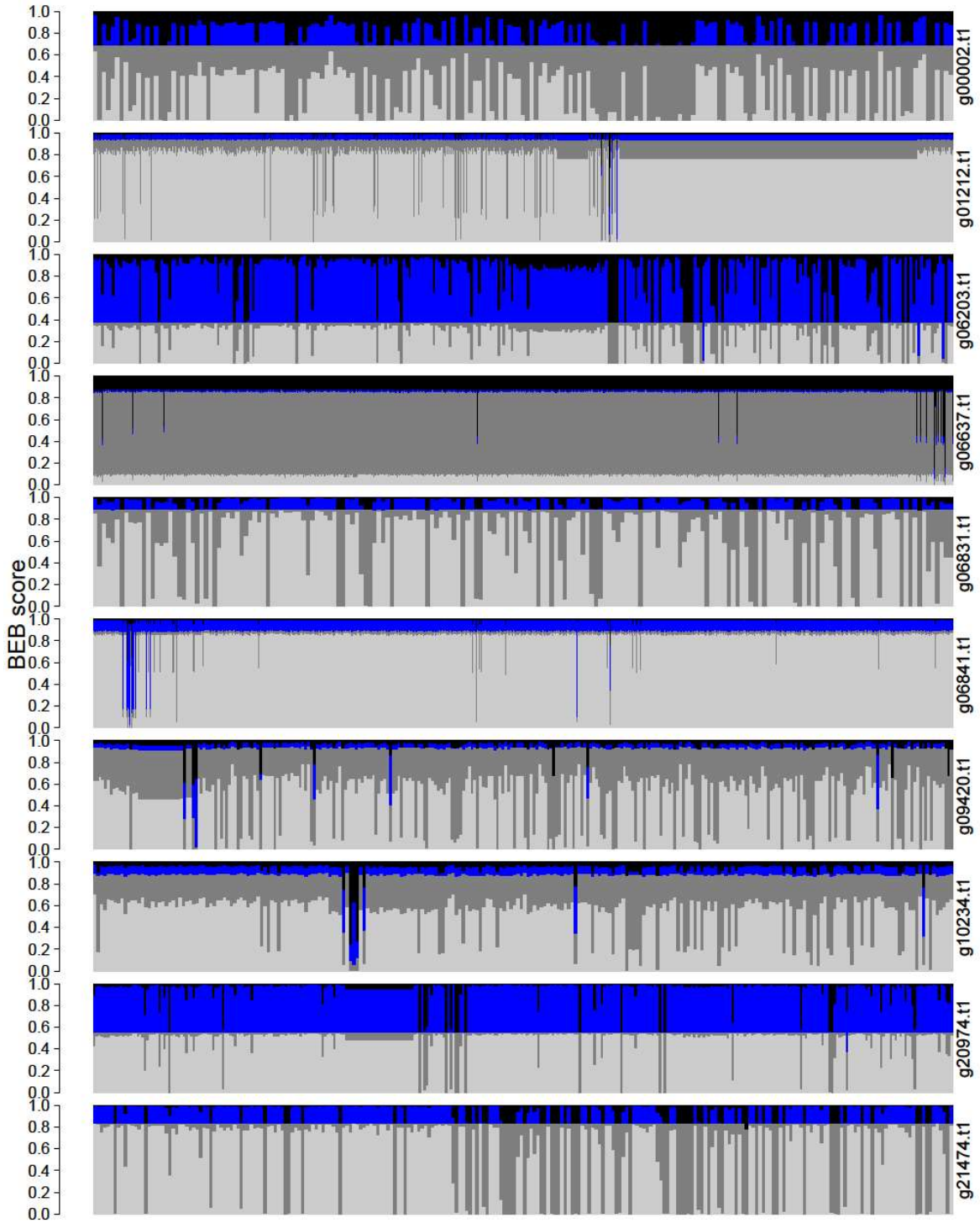


Fig A4.10. Bayesian probability of site class per codon, outputted by codeML branch site test, for genes outputted by codeML as having experienced episodic positive selection in the species C. pygargus*. x-axis: position along gene, y-axis:BEB-score. Lightgrey: $\omega_{for} < 1$ and $\omega_{back} < 1$. Darkgrey: $\omega_{for} = 1$ and $\omega_{back} = 1$. Red: $\omega_{for} > 1$ and $\omega_{back} < 1$. Black: $\omega_{for} > 1$ and $\omega_{back} = 1$. Note: for long genes (>2000 bp) colours of individual codons might be lost.*

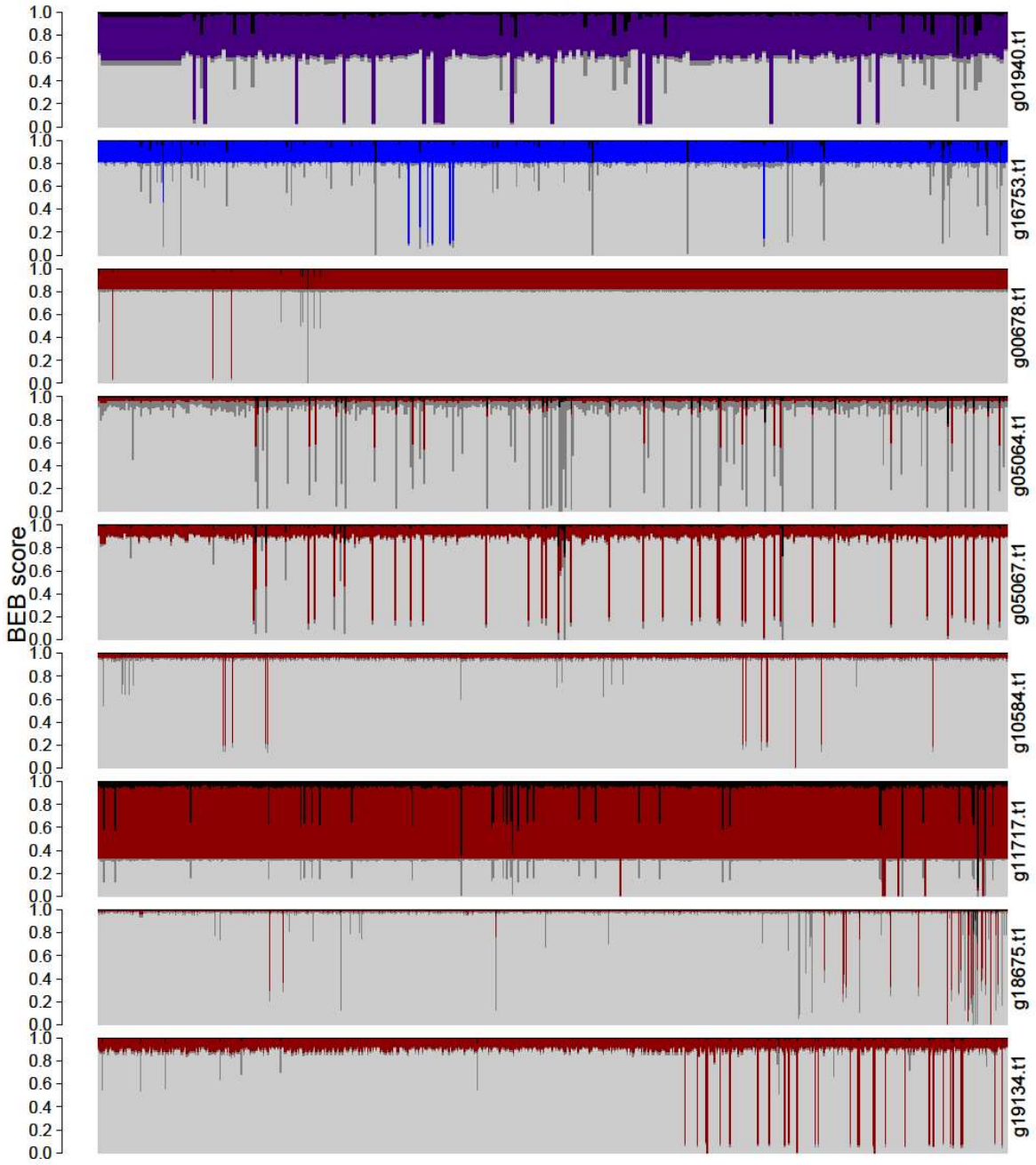


Fig A4.11. Bayesian probability of site classes per codon, outputted by codeML branch site test, for genes with accelerated dN/dS rates*. x-axis: position along gene, y-axis:BEB-score. Lightgrey: $\omega_{for} < 1$ and $\omega_{back} < 1$. Darkgrey: $\omega_{for} = 1$ and $\omega_{back} = 1$. Colour: $\omega_{for} > 1$ and $\omega_{back} < 1$ (purple: genus Capreolus as foreground lineage; blue: species C. pygargus as foreground lineage; red: species C. capreolus as foreground lineage). Black: $\omega_{for} > 1$ and $\omega_{back} = 1$. Note: for long genes (>2000 bp) colours of individual codons might be lost.*

Fig A4.12. Clusters of lineage specific amino acid mutations. *Multiple sequence alignments of a subset of genes marked by codeml branchsite tests as containing codons which have been under episodic positive selection in the species C. capreolus. These example genes illustrate that for most outlier genes, mutations (non-synonymous and synonymous alike) are clustered together, rather than spread throughout the gene, which is suggesting of a single genomic translocation event.*

Fig A4.12 cont. Clusters of lineage specific amino acid substitutions. *Multiple sequence alignments of a subset of genes marked by codeml branchsite tests as containing codons which have been under episodic positive selection in the species C. capreolus. These example genes illustrate that for most outlier genes,substitutions (non-synonymous and synonymous alike) are clustered together, rather than spread throughout the gene, which is suggesting of a single translocation event.*

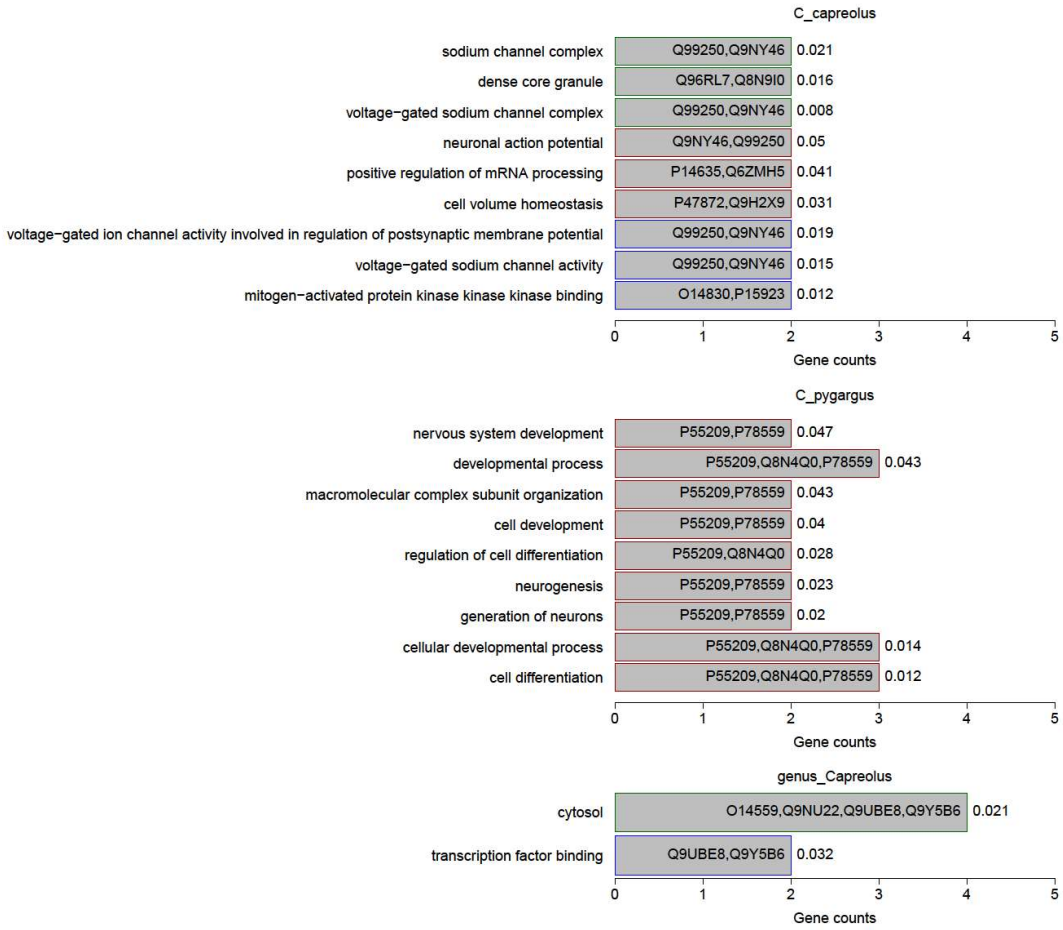


Fig A4.13A. GO enrichment analyses codeML outliers genes. *Enrichment GO accession terms for list of genes marked by codeML branchsite tests with respectively the species C. capreolus, C. pygargus and the genus Capreolus as foreground branches. Strings indicate Uniprot gene ID's, values indicated adjusted p-values. Colours indicate GO network categories: green = cellular component; red = biological process; blue = molecular function.*

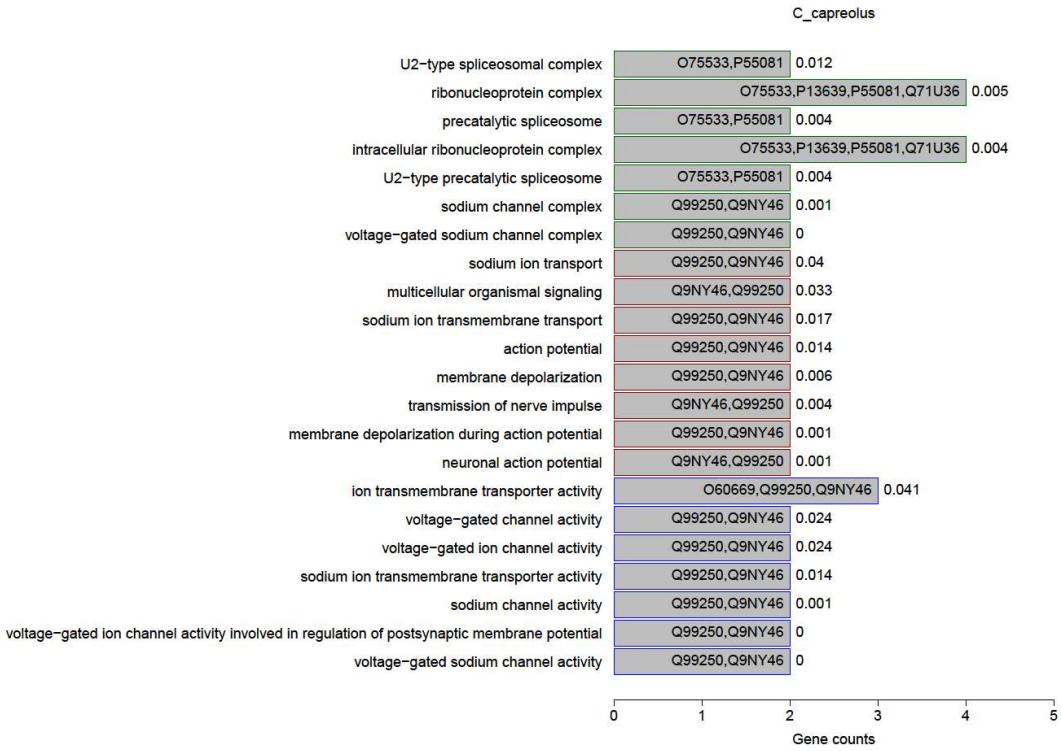


Fig A4.13B. GO enrichment analyses genes with accelerated dN/dS rates.
Enrichment GO accession terms for list of genes marked by the accelerated dN/dS tests as outlier genes. Strings indicate Uniprot gene ID's, values indicated adjusted p-values. Colours indicate GO network categories: green = cellular component; red = biological process; blue = molecular function. No significant results were found for C. pygargus and the genus Capreolus (as these outlier gene lists contained less than three genes.)

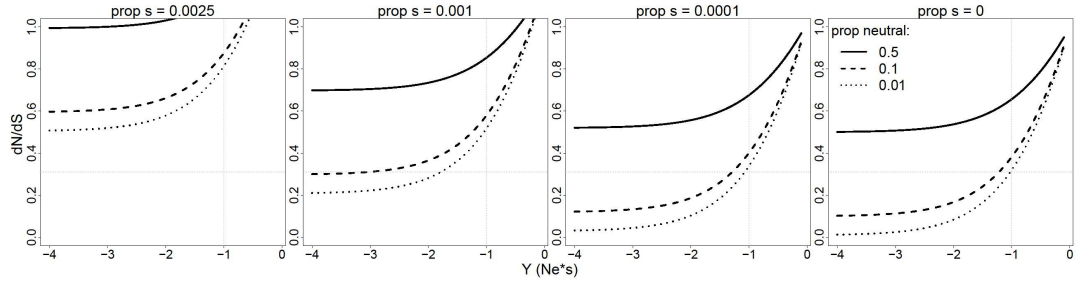


Fig A4.14. Expected average dN/dS values given various proportions of adaptive (s=0.01), neutral and deleterious mutations, and given a range of negative selection coefficients experienced by the deleterious mutations.

*Plots depicting the expected relation between dN/dS and the scaled selection coefficient Y (Ne*s), assuming that synonymous mutations are completely neutral (i.e. no codon usage bias), as described by the formula:*

dN/dS = f(N)/f(S)

*dN/dS = (prop(s)*f(s)+ prop(n)*f(n)+ prop(d)*f(d))/f(n)*

In which:

f(N): fixation probability of non-synonymous (N) mutations

f(S): fixation probability of synonymous (S) mutations = 1/N

prop(s,n,d): proportion N mutations which are positive selected (s), neutral (n) and deleterious (d)

f(s,n,d): fixation probability s, n and d mutations

If assuming that all N mutations are deleterious (i.e. s = 0, n = 0 and d = 1), the formula simplifies to (Kimura, 1962; Mugal et al., 2013):

dN/dS = f(d)/f(n)

$dN/dS = ((1-e^{-2s})/(1-e^{-2Ns}))/(1/N)$

$dN/dS = ((1-e^{-2Y/N})/(1-e^{-2Y}))/(1/N)$

$dN/dS \approx 1/N*2Y/(1-e^{-2Y}))/(1/N)$

$dN/dS \approx 2Y/(1-e^{-2Y})$

Vertical dashed lines indicate Y = -1 and the associated dN/dS-value given d=1.

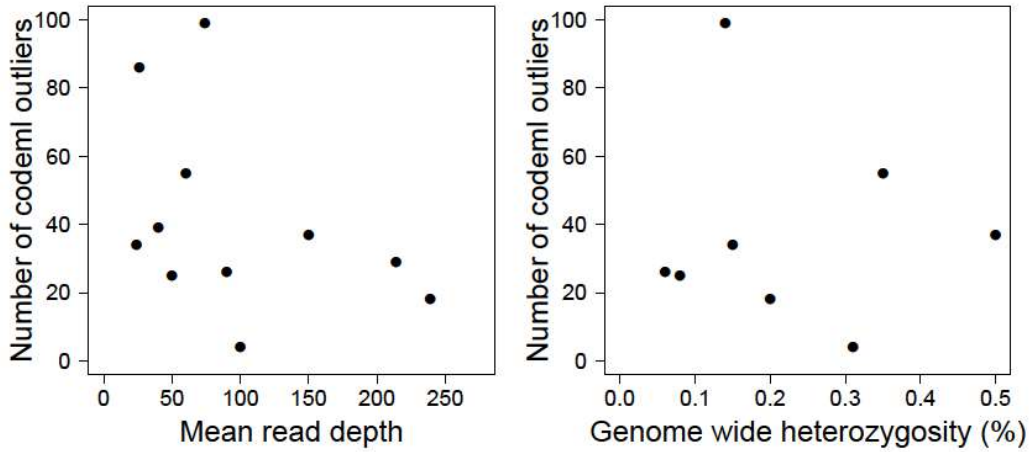


Figure A4.15. Number of codeml PSGs vs genome quality and genetic diversity. *Number of genes marked by codeml branch-site tests as putatively positively selected genes (PSGs) for various foreground branches (see Table A4.14), compared to genome quality (average genome wide read depth) and genome wide heterozygosity. Heterozygosity estimates were missing for various species, and hence the lower number of data points.*

References

Agaba, M., Ishengoma, E., Miller, W.C., McGrath, B.C., Hudson, C.N., Reina, O.C.B., Ratan, A., Burhans, R., Chikhi, R., Medvedev, P., et al. (2016). Giraffe genome sequence reveals clues to its unique morphology and physiology. Nat. Commun. *7*, 1–8.

Akashi, H. (1994). Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy. Genetics *136*, 927–935.

Akashi, H. (1995). Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in Drosophila DNA. Genetics *139*, 1067–1076.

Akey, J.M. (2009). Constructing genomic maps of positive selection in humans: where do we go from here? Genome Res. *19*, 711–722.

Akey, J.M., Zhang, G., Zhang, K., Jin, L., and Shriver, M.D. (2002). Interrogating a high-density SNP map for signatures of natural selection. Genome Res. *12*, 1805–1814.

Allendorf, F.W., and Lundquist, L.L. (2003). Introduction: Population Biology, Evolution, and Control of Invasive Species. Conserv. Biol. *17*, 24–30.

Andersen, R., Duncan, P., and Linnell, J.D.C. (1998). The European roe deer: the biology of success (Scandinavian University Press).

Andrew, S.C., Jensen, H., Hagen, I.J., Lundregan, S., and Griffith, S.C. (2018). Signatures of genetic adaptation to extremely varied Australian environments in introduced European house sparrows. Mol. Ecol. *27*, 4542–4555.

Antao, T., Lopes, A., Lopes, R.J., Beja-Pereira, A., and Luikart, G. (2008). LOSITAN: A workbench to detect molecular adaptation based on a Fst-outlier method. BMC Bioinformatics *9*, 323.

Atterby, H., Allnutt, T.R., MacNicoll, A.D., Jones, E.P., and Smith, G.C. (2015). Population genetic structure of the red fox (<Emphasis Type="Italic">Vulpes vulpes</Emphasis>) in the UK. Mammal Res. *60*, 9–19.

Avise, J.C. (2000). Phylogeography: The History and Formation of Species (Harvard University Press).

Avise, J.C., Walker, D., and Johns, G.C. (1998). Speciation durations and Pleistocene effects on vertebrate phylogeography. Proc. R. Soc. B Biol. Sci. *265*, 1707–1712.

Ayala, F.J. (2000). Neutralism and selectionism: the molecular clock. Gene *261*, 27–33.

Baalsrud, H.T., Tørresen, O.K., Solbakken, M.H., Salzburger, W., Hanel, R., Jakobsen, K.S., and Jentoft, S. (2018). De Novo Gene Evolution of Antifreeze Glycoproteins in Codfishes Revealed by Whole Genome Sequence Data. Mol. Biol. Evol. *35*, 593–606.

Backman, T.W.H., and Girke, T. (2016). systemPipeR: NGS workflow and report generation environment. BMC Bioinformatics *17*, 388.

Baird, N.A., Etter, P.D., Atwood, T.S., Currey, M.C., Shiver, A.L., Lewis, Z.A., Selker, E.U., Cresko, W.A., and Johnson, E.A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. PloS One *3*, e3376.

Baker, K.H., and Hoelzel, A.R. (2014). Influence of Holocene environmental change and anthropogenic impact on the diversity and distribution of roe deer. Heredity *112*, 607–615.

Baker, K.H., and Rus Hoelzel, A. (2012). Evolution of population genetic structure of the British roe deer by natural and anthropogenic processes (*Capreolus capreolus*). Ecol. Evol. *3*, 89–102.

Bamshad, M., and Wooding, S.P. (2003). Signatures of natural selection in the human genome. Nat. Rev. Genet. *4*, 99–111.

Bana, N.Á., Nyiri, A., Nagy, J., Frank, K., Nagy, T., Stéger, V., Schiller, M., Lakatos, P., Sugár, L., Horn, P., et al. (2018). The red deer Cervus elaphus genome CerEla1.0: sequencing, annotating, genes, and chromosomes. Mol. Genet. Genomics MGG *293*, 665–684.

Baptestini, E.M., de Aguiar, M.A.M., and Bar-Yam, Y. (2013). Conditions for neutral speciation via isolation by distance. J. Theor. Biol. *335*, 51–56.

Barnosky, A.D. (2005). Effects of Quaternary Climatic Change on Speciation in Mammals. J. Mamm. Evol. *12*, 247–264.

Barraclough, T.G., and Vogler, A.P. (2000). Detecting the Geographical Pattern of Speciation from Species-Level Phylogenies. Am. Nat. *155*, 419–434.

Barrett, R.D.H., and Schluter, D. (2008). Adaptation from standing genetic variation. Trends Ecol. Evol. *23*, 38–44.

Barton, N.H., and Charlesworth, B. (1984). Genetic Revolutions, Founder Effects, and Speciation. Annu. Rev. Ecol. Syst. *15*, 133–164.

Batbaabtar, J., Gillespie, A.R., Fink, D., Matmon, A., and Fujioka, T. (2018). Asynchronous glaciations in arid continental climate. Quat. Sci. Rev. *182*, 1–19.

Beaumont, M.A. (2005). Adaptation and speciation: what can F(st) tell us? Trends Ecol. Evol. *20*, 435–440.

Beaumont, M.A., and Balding, D.J. (2004). Identifying adaptive genetic divergence among populations from genome scans. Mol. Ecol. *13*, 969–980.

Beaumont, M.A., and Nichols, R.A. (1996). Evaluating Loci for Use in the Genetic Analysis of Population Structure. Proc. R. Soc. Lond. B Biol. Sci. *263*, 1619–1626.

Bedford, T., and Hartl, D.L. (2008). Overdispersion of the molecular clock: temporal variation of gene-specific substitution rates in Drosophila. Mol. Biol. Evol. *25*, 1631–1638.

Begun, D.J., and Aquadro, C.F. (1992). Levels of naturally occurring DNA polymorphism correlate with recombination rates in D. melanogaster. Nature *356*, 519–520.

Begun, D.J., Holloway, A.K., Stevens, K., Hillier, L.W., Poh, Y.-P., Hahn, M.W., Nista, P.M., Jones, C.D., Kern, A.D., Dewey, C.N., et al. (2007). Population genomics: whole-genome analysis of polymorphism and divergence in Drosophila simulans. PLoS Biol. *5*, e310.

Beichman, A.C., Koepfli, K.-P., Li, G., Murphy, W., Dobrynin, P., Kilver, S., Tinker, M.T., Murray, M.J., Johnson, J., Lindblad-Toh, K., Karlsson, E. K., Lohmueller, K. E., Wayne, R. K. (2019). Aquatic adaptation and depleted diversity: a deep dive into the genomes of the sea otter and giant otter. Mol. Biol. Evol.

Bell, R.C., Brasileiro, C.A., Haddad, C.F.B., and Zamudio, K.R. (2012). Evolutionary history of Scinax treefrogs on land-bridge islands in south-eastern Brazil. J. Biogeogr. *39*, 1733–1742.

Benton, M.J. (2009). The Red Queen and the Court Jester: species diversity and the role of biotic and abiotic factors through time. Science *323*, 728–732.

Bernardi, G., Azzurro, E., Golani, D., and Miller, M.R. (2016). Genomic signatures of rapid adaptive evolution in the bluespotted cornetfish, a Mediterranean Lessepsian invader. Mol. Ecol. *25*, 3384–3396.

Bernatchez, S., Laporte, M., Perrier, C., Sirois, P., and Bernatchez, L. (2016). Investigating genomic and phenotypic parallelism between piscivorous and planktivorous lake trout (Salvelinus namaycush) ecotypes by means of RADseq and morphometrics analyses. Mol. Ecol. *25*, 4773–4792.

Besenbacher, S., Sulem, P., Helgason, A., Helgason, H., Kristjansson, H., Jonasdottir, A., Jonasdottir, A., Magnusson, O.T., Thorsteinsdottir, U., Masson, G., et al. (2016). Multi-nucleotide de novo Mutations in Humans. PLoS Genet. *12*, e1006315.

Bibi, F., and Kiessling, W. (2015). Continuous evolutionary change in Plio-Pleistocene mammals of eastern Africa. Proc. Natl. Acad. Sci. *112*, 10623–10628.

Birky, C.W., and Walsh, J.B. (1988). Effects of linkage on rates of molecular evolution. Proc. Natl. Acad. Sci. U. S. A. *85*, 6414–6418.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120.

Bosse, M., Spurgin, L.G., Laine, V.N., Cole, E.F., Firth, J.A., Gienapp, P., Gosler, A.G., McMahon, K., Poissant, J., Verhagen, I., et al. (2017). Recent natural selection causes adaptive evolution of an avian polygenic trait. Science *358*, 365–368.

Bourret, V., Kent, M.P., Primmer, C.R., Vasemägi, A., Karlsson, S., Hindar, K., McGinnity, P., Verspoor, E., Bernatchez, L., and Lien, S. (2013). SNP-array reveals genome-wide patterns of geographical and potential adaptive divergence across the natural range of Atlantic salmon (Salmo salar). Mol. Ecol. *22*, 532–551.

Bovine Genome Sequencing and Analysis Consortium, Elsik, C.G., Tellam, R.L., Worley, K.C., Gibbs, R.A., Muzny, D.M., Weinstock, G.M., Adelson, D.L., Eichler, E.E., Elnitski, L., et al. (2009). The genome sequence of taurine cattle: a window to ruminant biology and evolution. Science *324*, 522–528.

Bradshaw, W.E., and Holzapfel, C.M. (2010). Light, time, and the physiology of biotic response to rapid climate change in animals. Annu. Rev. Physiol. *72*, 147–166.

Brakefield, P.M., and de Jong, P.W. (2011). A steep cline in ladybird melanism has decayed over 25 years: a genetic response to climate change? Heredity *107*, 574–578.

Brandrud, M.K., Paun, O., Lorenzo, M.T., Nordal, I., and Brysting, A.K. (2017). RADseq provides evidence for parallel ecotypic divergence in the autotetraploid Cochlearia officinalis in Northern Norway. Sci. Rep. *7*, 5573.

Brawand, D., Wagner, C.E., Li, Y.I., Malinsky, M., Keller, I., Fan, S., Simakov, O., Ng, A.Y., Lim, Z.W., Bezault, E., et al. (2014). The genomic substrate for adaptive radiation in African cichlid fish. Nature *513*, 375–381.

Britton, T., Anderson, C.L., Jacquet, D., Lundqvist, S., and Bremer, K. (2007). Estimating divergence times in large phylogenetic trees. Systematic Biology. *56(5)*, 741-752.

BROWN, O.J.F. (2006). Tasmanian devil (Sarcophilus harrisii) extinction on the Australian mainland in the mid-Holocene: multicausality and ENSO intensification. Alcheringa Australas. J. Palaeontol. *30*, 49–57.

Brues, A.M. (1964). The cost of evolution vs. the cost of not evolving. Evolution *18*, 379–383.

Brüniche-Olsen, A., Kellner, K. F., Anderson, C. J., De Woody, J. A. (2018). Runs of homozygosity have utility in mammalian conservation and evolutionary studies. Conservation Genetics. *19*,1295-1307.

Burridge, C.P., Brown, W.E., Wadley, J., Nankervis, D.L., Olivier, L., Gardner, M.G., Hull, C., Barbour, R., and Austin, J.J. (2013). Did postglacial sea-level changes initiate the evolutionary divergence of a Tasmanian endemic raptor from its mainland relative? Proc. Biol. Sci. *280*, 20132448.

Cammen, K.M., Schultz, T.F., Rosel, P.E., Wells, R.S., and Read, A.J. (2015). Genomewide investigation of adaptation to harmful algal blooms in common bottlenose dolphins (Tursiops truncatus). Mol. Ecol. *24*, 4697–4710.

Carroll, S.P., Hendry, A.P., Reznick, D.N., and Fox, C.W. (2007). Evolution on ecological time-scales. Funct. Ecol. *21*, 387–393.

Catchen, J., Hohenlohe, P.A., Bassham, S., Amores, A., and Cresko, W.A. (2013). Stacks: an analysis tool set for population genomics. Mol. Ecol. *22*, 3124–3140.

Chadarevian, S. de. 1999. Protein sequencing and the making of molecular genetics. Trends Biochem Sci. *24(5*), 203-6

Charlesworth, B. (2012). The effects of deleterious mutations on evolution at linked sites. Genetics *190*, 5–22.

Charlesworth, J., and Eyre-Walker, A. (2008). The McDonald-Kreitman test and slightly deleterious mutations. Mol. Biol. Evol. *25*, 1007–1015.

Charlesworth, B., Morgan, M.T., and Charlesworth, D. (1993). The effect of deleterious mutations on neutral molecular variation. Genetics *134*, 1289–1303.

Chen, L., Qiu, Q., Jiang, Y., Wang, K., Lin, Z., Li, Z., Bibi, F., Yang, Y., Wang, J., Nie, W., et al. (2019). Large-scale ruminant genome sequencing provides insights into their evolution and distinct traits. Science *364*.

Chen, Z., Farrell, A.P., Matala, A., Hoffman, N., and Narum, S.R. (2018). Physiological and genomic signatures of evolutionary thermal adaptation in redband trout from extreme climates. Evol. Appl. *11*, 1686–1699.

Cho, Y.S., Hu, L., Hou, H., Lee, H., Xu, J., Kwon, S., Oh, S., Kim, H.-M., Jho, S., Kim, S., et al. (2013). The tiger genome and comparative analysis with lion and snow leopard genomes. Nat. Commun. *4*, 1–7.

Colautti, R.I., and Lau, J.A. (2015). Contemporary evolution during invasion: evidence for differentiation, natural selection, and local adaptation. Mol. Ecol. *24*, 1999–2017.

Coles, B.J. (1998). Doggerland: a Speculative Survey. Proc. Prehist. Soc. *64*, 45–81.

Comeron, J.M. (2017). Background selection as null hypothesis in population genomics: insights and challenges from Drosophila studies. Philos. Trans. R. Soc. Lond. B. Biol. Sci. *372*.

Comes, H.P., Tribsch, A., and Bittkau, C. (2008). Plant speciation in continental island floras as exemplified by Nigella in the Aegean Archipelago. Philos. Trans. R. Soc. Lond. B. Biol. Sci. *363*, 3083–3096.

Cook, L. M., and Saccheri, I. J. (2013). The peppered moth and industrial melanism: evolution of a natural selection case study. Heredity. *110,* 207-212.

Cooper, G. M., Stone, E. A., Asimenos, G., NISC Comparative Sequencing Program, Green, E. D., Batzoglou, S., Sidow, A. (2005). Distribution and intensity of constraint in mammalian genomic sequence. Genome Research. *15(7)*, 901-913.

Corbett-Detig, R.B., Hartl, D.L., and Sackton, T.B. (2015). Natural selection constrains neutral diversity across a wide range of species. PLoS Biol. *13*, e1002112.

Curnoe, D., Thorne, A., and Coate, J.A. (2006). Timing and tempo of primate speciation. J. Evol. Biol. *19*, 59–65.

Cutler, D.J. (2000). Understanding the overdispersed molecular clock. Genetics *154*(3), 1403-17.

Cutter, A.D., and Payseur, B.A. (2013). Genomic signatures of selection at linked sites: unifying the disparity among species. Nat. Rev. Genet. *14*, 262–274.

Czekanski-Moir, J.E., and Rundell, R.J. (2019). The Ecology of Nonecological Speciation and Nonadaptive Radiations. Trends Ecol. Evol. *34*, 400–415.

Dalongeville, A., Benestan, L., Mouillot, D., Lobreaux, S., and Manel, S. (2018). Combining six genome scan methods to detect candidate genes to salinity in the Mediterranean striped red mullet (Mullus surmuletus). BMC Genomics *19*.

Danilkin, A. (1995). Behavioural Ecology of Siberian and European Roe Deer.

Darcey, J., Horner, K., Walsh, T., Southern, H., Marjanovic, E.J., and Devlin, H. (2013). Tooth loss and osteoporosis: to assess the association between osteoporosis status and tooth number. Br. Dent. J. *214*, E10.

Darwin, C., and Wallace, A. (1858). On the Tendency of Species to form Varieties; and on the Perpetuation of Varieties and Species by Natural Means of Selection. J. Proc. Linn. Soc. Lond. Zool. *3*, 45–62.

Davydov, E. V., Goode, D. L., Sirota, M., Cooper, G. M., Sidow, A., Batzoglou, S. (2010). Identifying a high fraction of the human genome to be under selective constraint using GERP++. Plos Computational Biology.

De Mita, S., Thuillet, A.-C., Gay, L., Ahmadi, N., Manel, S., Ronfort, J., and Vigouroux, Y. (2013). Detecting selection along environmental gradients: analysis of eight methods and their effectiveness for outbreeding and selfing populations. Mol. Ecol. *22*, 1383–1399.

Delsuc, F., and Tilak, M.-K. (2015). Naked but not Hairless: the pitfalls of analyses of molecular adaptation based on few genome sequence comparisons. Genome Biol. Evol. *7*, 768–774.

Dickerson, R.E. (1971). The structures of cytochrome c and the rates of molecular evolution. J. Mol. Evol. *1*, 26–45.

Diekmann, Y., and Pereira-Leal, J.B. (2016). Gene Tree Affects Inference of Sites Under Selection by the Branch-Site Test of Positive Selection. Evol. Bioinforma. Online *11*, 11–17.

Diller, K.C., Gilbert, W.A., and Kocher, T.D. (2002). Selective sweeps in the human genome: a starting point for identifying genetic differences between modern humans and chimpanzees. Mol. Biol. Evol. *19*, 2342–2345.

Dodson, E. (1962). Note on the cost of natural selection.

Douzery, E., and Randi, E. (1997). The mitochondrial control region of Cervidae: evolutionary patterns and phylogenetic content. Mol. Biol. Evol. *14*, 1154–1166.

Drummond, A.J., Ho, S.Y.W., Phillips, M.J., and Rambaut, A. (2006). Relaxed phylogenetics and dating with confidence. PLoS Biol. *4*, e88.

Duforet-Frebourg, N., Bazin, E., and Blum, M.G.B. (2014). Genome scans for detecting footprints of local adaptation using a Bayesian factor model. Mol. Biol. Evol. *31*, 2483–2495.

Durinck, S., Spellman, P.T., Birney, E., and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat. Protoc. *4*, 1184–1191.

Eckert, C.G., Samis, K.E., and Lougheed, S.C. (2008). Genetic variation across species' geographical ranges: the central-marginal hypothesis and beyond. Mol. Ecol. *17*, 1170–1188.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. *32*, 1792–1797.

Edwards, C.J., Soulsbury, C.D., Statham, M.J., Ho, S.Y.W., Wall, D., Dolf, G., Iossa, G., Baker, P.J., Harris, S., Sacks, B.N., et al. (2012). Temporal genetic variation of the red fox, Vulpes vulpes, across western Europe and the British Isles. Quat. Sci. Rev. *57*, 95–104.

Endler, J. (1977). Geographic Variation, Speciation and Clines (Princeton, NJ: Princeton University Press).

Endler, J. (1986). Natural Selection in the Wild.

Excoffier, L., and Lischer, H.E.L. (2010). Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Mol. Ecol. Resour. *10*, 564–567.

Eyre-Walker, A., and Keightley, P.D. (2007). The distribution of fitness effects of new mutations. Nat. Rev. Genet. *8*, 610–618.

Eyre-Walker, A., and Keightley, P.D. (2009). Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. Mol. Biol. Evol. *26*, 2097–2108.

Eyre-Walker, A., Woolfit, M., and Phelps, T. (2006). The distribution of fitness effects of new deleterious amino acid mutations in humans. Genetics *173*, 891–900.

Fay, J.C. (2011). Weighing the evidence for adaptation at the molecular level. Trends Genet. TIG *27*, 343–349.

Fay, J.C., and Wu, C.I. (2000). Hitchhiking under positive Darwinian selection. Genetics *155*, 1405–1413.

Fay, J.C., Wyckoff, G.J., and Wu, C.I. (2001). Positive and negative selection on the human genome. Genetics *158*, 1227–1234.

Felsenstein, J. (1971). On the Biological Significance of the Cost of Gene Substitution. Am. Nat. *105*, 1–11.

Feng, S., Fang, Q., Barnett, R., Li, C., Han, S., Kuhlwilm, M., Zhou, L., Pan, H., Deng, Y., Chen, G., et al. (2019). The Genomic Footprints of the Fall and Recovery of the Crested Ibis. Curr. Biol. CB *29*, 340–349.e7.

Figueiró, H.V., Li, G., Trindade, F.J., Assis, J., Pais, F., Fernandes, G., Santos, S.H.D., Hughes, G.M., Komissarov, A., Antunes, A., et al. (2017). Genome-wide signatures of complex introgression and adaptive evolution in the big cats. Sci. Adv. *3*, e1700299.

Flanagan, S.P., and Jones, A.G. (2017). Constraints on the FST-Heterozygosity Outlier Approach. J. Hered. *108*, 561–573.

Fletcher, W., and Yang, Z. (2010). The effects of insertions, deletions and alignment errors on the branch-site test of positive selection. Molecular Biology and Evolution. *27(10)*, 2257-2267.

Foll, M., and Gaggiotti, O. (2008). A Genome-Scan Method to Identify Selected Loci Appropriate for Both Dominant and Codominant Markers: A Bayesian Perspective. Genetics *180*, 977–993.

Foote, A.D., Liu, Y., Thomas, G.W.C., Vinař, T., Alföldi, J., Deng, J., Dugan, S., van Elk, C.E., Hunter, M.E., Joshi, V., et al. (2015). Convergent evolution of the genomes of marine mammals. Nat. Genet. *47*, 272–275.

Foster, J.B. (1964). Evolution of Mammals on Islands. Nature *202*, 234–235.

Frantz, A.C., McDevitt, A.D., Pope, L.C., Kochan, J., Davison, J., Clements, C.F., Elmeros, M., Molina-Vacas, G., Ruiz-Gonzalez, A., Balestrieri, A., et al. (2014). Revisiting the phylogeography and demography of European badgers (Meles meles) based on broad sampling, multiple markers and simulations. Heredity *113*, 443–453.

Frantz, L.A.F., Madsen, O., Megens, H.-J., Schraiber, J.G., Paudel, Y., Bosse, M., Crooijmans, R.P.M.A., Larson, G., and Groenen, M.A.M. (2015). Evolution of Tibetan Wild Boars. Nat. Genet. *47*, 188–189.

Frichot, E., and François, O. (2015). LEA: An R package for landscape and ecological association studies. Methods Ecol. Evol. *6*, 925–929.

Fu, Y.X., and Li, W.H. (1993). Statistical tests of neutrality of mutations. Genetics *133*, 693–709.

Funk, W.C., Lovich, R.E., Hohenlohe, P.A., Hofman, C.A., Morrison, S.A., Sillett, T.S., Ghalambor, C.K., Maldonado, J.E., Rick, T.C., Day, M.D., et al. (2016). Adaptive divergence despite strong genetic drift: genomic analysis of the evolutionary mechanisms causing genetic differentiation in the island fox (Urocyon littoralis). Mol. Ecol. *25*, 2176–2194.

Gaither, M.R., Bernal, M.A., Coleman, R.R., Bowen, B.W., Jones, S.A., Simison, W.B., and Rocha, L.A. (2015). Genomic signatures of geographic isolation and natural selection in coral reef fishes. Mol. Ecol. *24*, 1543–1557.

Galtier, N. (2016). Adaptive Protein Evolution in Animals and the Effective Population Size Hypothesis. PLoS Genet. *12*.

Gautier, M., Moazami-Goudarzi, K., Levéziel, H., Parinello, H., Grohs, C., Rialle, S., Kowalczyk, R., and Flori, L. (2016). Deciphering the Wisent Demographic and Adaptive Histories from Individual Whole-Genome Sequences. Mol. Biol. Evol. *33*, 2801–2814.

Gervais, L., Perrier, C., Bernard, M., Merlet, J., Pemberton, J.M., Pujol, B., and Quéméré, E. (2019). RAD-sequencing for estimating genomic relatedness matrix-based heritability in the wild: A case study in roe deer. Mol. Ecol. Resour. *19*, 1205–1217.

Gharib, W.H., and Robinson-Rechavi, M. (2013). The branch-site test of positive selection is surprisingly robust but lacks power under synonymous substitution saturation and variation in GC. Mol. Biol. Evol. *30*, 1675–1686.

Gillespie, J.H. (1989). Lineage effects and the index of dispersion of molecular evolution. Mol. Biol. Evol. *6*, 636–647.

Gillespie, J.H. (2000). Genetic drift in an infinite population. The pseudohitchhiking model. Genetics *155*, 909–919.

Gittenberger, E. (1991). What about non-adaptive radiation? Biol. J. Linn. Soc. *43*, 263–272.

Gossmann, T.I., Song, B.-H., Windsor, A.J., Mitchell-Olds, T., Dixon, C.J., Kapralov, M.V., Filatov, D.A., and Eyre-Walker, A. (2010). Genome Wide Analyses Reveal Little Evidence for Adaptive Evolution in Many Plant Species. Mol. Biol. Evol. *27*, 1822–1832.

Gossmann, T.I., Keightley, P.D., and Eyre-Walker, A. (2012). The effect of variation in the effective population size on the rate of adaptive molecular evolution in eukaryotes. Genome Biol. Evol. *4*, 658–667.

Gower (2019). Inferring the characteristics of ancient populations using bioinformation analyses of genome-wide DNA sequencing data.

Gray, M.M., Parmenter, M.D., Hogan, C.A., Ford, I., Cuthbert, R.J., Ryan, P.G., Broman, K.W., and Payseur, B.A. (2015). Genetics of Rapid and Extreme Size Evolution in Island Mice. Genetics *201*, 213–228.

Green, R.E., Braun, E.L., Armstrong, J., Earl, D., Nguyen, N., Hickey, G., Vandewege, M.W., St John, J.A., Capella-Gutiérrez, S., Castoe, T.A., et al. (2014). Three crocodilian genomes reveal ancestral patterns of evolution among archosaurs. Science *346*, 1254449.

Groenen, M.A.M., Archibald, A.L., Uenishi, H., Tuggle, C.K., Takeuchi, Y., Rothschild, M.F., Rogel-Gaillard, C., Park, C., Milan, D., Megens, H.-J., et al. (2012). Analyses of pig genomes provide insight into porcine demography and evolution. Nature *491*, 393–398.

Guo, Z., Neilson, L.J., Zhong, H., Murray, P.S., Zanivan, S., and Zaidel-Bar, R. (2014). E-cadherin interactome complexity and robustness resolved by quantitative proteomics. Sci. Signal. *7*, rs7.

Haasl, R.J., and Payseur, B.A. (2016). DETECTING SELECTION IN NATURAL POPULATIONS: MAKING SENSE OF GENOME SCANS AND TOWARDS ALTERNATIVE SOLUTIONS. Mol. Ecol. *25*, 5–23.

Haffer, J. (1969). Speciation in amazonian forest birds. Science *165*, 131–137.

Hahn, M.W. (2008). Toward a selection theory of molecular evolution. Evol. Int. J. Org. Evol. *62*, 255–265.

Haldane, J.B.S. (1957). The cost of natural selection. J. Genet. *55*, 511.

Haller, B., C. and Messer, P. W. (2019). SLIM 3: Forward genetic simulations beyong the Wright-Fisher model. *36(3)*, 632-637.

Harris, H. (1966). Enzyme polymorphisms in man. Proc. R. Soc. Lond. B. Biol. Sci. *164(995)*: 298-310.

Harris, R.S. Improved pairwise alignment of genomic dna.

Hartl, D.L., and Clark, A.G. (1997). Principles of Population Genetics (Sinauer Associates).

Hassett, J., Browne, K.A., McCormack, G.P., Moore, E., Society, N.I.H.B., Soland, G., and Geary, M. (2018). A significant pure population of the dark European honey bee (Apis mellifera mellifera) remains in Ireland.

Hawks, J., Wang, E.T., Cochran, G.M., Harpending, H.C., and Moyzis, R.K. (2007). Recent acceleration of human adaptive evolution. Proc. Natl. Acad. Sci. U. S. A. *104*, 20753–20758.

He, Z., Chen, Q., Yang, H., Chen, Q., Shi, S., and Wu, C.-I. (2018). Conflicting signals of adaptive molecular evolution - Where does the neutral theory stand after 50 years? bioRxiv 417717.

Hendrickson, S.L. (2013). A genome wide study of genetic adaptation to high altitude in feral Andean Horses of the páramo. BMC Evol. Biol. *13*, 273.

Hendry, A.P., and Kinnison, M.T. (1999). Perspective: The Pace of Modern Life: Measuring Rates of Contemporary Microevolution. Evolution *53*, 1637–1653.

Hendry, A.P., Nosil, P., and Rieseberg, L.H. (2007). The speed of ecological speciation. Funct. Ecol. *21*, 455–464.

Heppenheimer, E., Brzeski, K.E., Hinton, J.W., Patterson, B.R., Rutledge, L.Y., DeCandia, A.L., Wheeldon, T., Fain, S.R., Hohenlohe, P.A., Kays, R., et al. (2018). High genomic diversity and candidate genes under selection associated with range expansion in eastern coyote (Canis latrans) populations. Ecol. Evol. *8*, 12641–12655.

Hermisson, J., and Pennings, P.S. (2005). Soft sweeps: molecular population genetics of adaptation from standing genetic variation. Genetics *169*, 2335–2352.

Hermisson, J., and Pennings, P.S. (2017). Soft sweeps and beyond: understanding the patterns and probabilities of selection footprints under rapid adaptation. Methods Ecol. Evol. *8*, 700–716.

Hess, J.E., Campbell, N.R., Close, D.A., Docker, M.F., and Narum, S.R. (2013). Population genomics of Pacific lamprey: adaptive variation in a highly dispersive species. Mol. Ecol. *22*, 2898–2916.

Hewitt, G. (2000). The genetic legacy of the Quaternary ice ages. Nature *405*, 907–913.

Hewitt, G.M. (1999). Post-glacial re-colonization of European biota. Biol. J. Linn. Soc. *68*, 87–112.

Hewitt, G.M. (2004). Genetic consequences of climatic oscillations in the Quaternary. Philos. Trans. R. Soc. B Biol. Sci. *359*, 183–195.

Heymann, R., About, I., Lendahl, U., Franquin, J.-C., Öbrink, B., and Mitsiadis, T.A. (2002). E- and N-Cadherin Distribution in Developing and Functional Human Teeth under Normal and Pathological Conditions. Am. J. Pathol. *160*, 2123–2133.

Hickey, D., and Golding, B. (2019). Sex Solves Haldane's Dilemma. Genome.

Hof, A.E. van't, Campagne, P., Rigden, D.J., Yung, C.J., Lingley, J., Quail, M.A., Hall, N., Darby, A.C., and Saccheri, I.J. (2016). The industrial melanism mutation in British peppered moths is a transposable element. Nature *534*, 102–105.

Hofman, C.A., Rick, T.C., Hawkins, M.T.R., Funk, W.C., Ralls, K., Boser, C.L., Collins, P.W., Coonan, T., King, J.L., Morrison, S.A., et al. (2015). Mitochondrial Genomes Suggest Rapid Evolution of Dwarf California Channel Islands Foxes (Urocyon littoralis). PLoS ONE *10*.

Hofmann, R.R. (1989). Evolutionary steps of ecophysiological adaptation and diversification of ruminants: a comparative view of their digestive system. Oecologia *78*, 443–457.

Hofreiter, M., and Stewart, J. (2009). Ecological change, range fluctuations and population dynamics during the Pleistocene. Curr. Biol. CB *19*, R584-594.

Hohenlohe, P.A., Bassham, S., Etter, P.D., Stiffler, N., Johnson, E.A., and Cresko, W.A. (2010). Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. PLoS Genet. *6*, e1000862.

Hoskin, C.J., Higgie, M., McDonald, K.R., and Moritz, C. (2005). Reinforcement drives rapid allopatric speciation. Nature *437*, 1353–1356.

Hudson, R.R. (1982). Testing the constant-rate neutral allele model with protein sequence data. Evolution *37*, 203-217.

Hudson, R.R., Kreitman, M., and Aguadé, M. (1987). A test of neutral molecular evolution based on nucleotide data. Genetics *116*, 153–159.

Hughes, A.L. (2007). Looking for Darwin in all the wrong places: the misguided quest for positive selection at the nucleotide sequence level. Heredity *99*, 364–373.

Hughes, A.L., and Nei, M. (1988). Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature *335*, 167–170.

Hughes, A.L., Packer, B., Welch, R., Bergen, A.W., Chanock, S.J., and Yeager, M. (2003). Widespread purifying selection at polymorphic sites in human protein-coding loci. Proc. Natl. Acad. Sci. U. S. A. *100*, 15754–15757.

Hurston, H., Voith, L., Bonanno, J., Foufopoulos, J., Pafilis, P., Valakos, E., and Anthony, N. (2009). Effects of fragmentation on genetic diversity in island populations of the Aegean wall lizard Podarcis erhardii (Lacertidae, Reptilia). Mol. Phylogenet. Evol. *52*, 395–405.

Itescu, Y., Foufopoulos, J., Pafilis, P., and Meiri, S. (2019). The diverse nature of island isolation and its effect on land bridge insular faunas. Glob. Ecol. Biogeogr. *n/a*.

Janecka, J., Chowdhary, B., and Murphy, W. (2012). Exploring the correlations between sequence evolution rate and phenotypic divergence across the Mammalian tree provides insights into adaptive evolution. J. Biosci. *37*, 897–909.

Jenkins, D.L., Ortori, C.A., and Brookfield, J.F. (1995). A test for adaptive change in DNA sequences controlling transcription. Proc. Biol. Sci. *261*, 203–207.

Jensen, J.D., Payseur, B.A., Stephan, W., Aquadro, C.F., Lynch, M., Charlesworth, D., and Charlesworth, B. (2019). The importance of the Neutral Theory in 1968 and 50 years on: A response to Kern and Hahn 2018. Evol. Int. J. Org. Evol. *73*, 111–114.

Johnston, R.F., and Selander, R.K. (1964). House Sparrows: Rapid Evolution of Races in North America. Science *144*, 548–550.

Jombart, T. (2008). Adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics *24*, 1403–1405.

Jombart, T., and Ahmed, I. (2011). Adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. Bioinformatics *27*, 3070–3071.

Jordan, G., and Goldman, N. (2012). The effects of alignment error and alignment filtering on the sitewise detection of positive selection. Molecular Biology and Evolution. *29(4), 1125-1139*.

Harrison, P. W., Jordan, G. E., Montgomery, S. H. (2014). SWAMP: sliding window alignment masker for PAML. Evol. Bioinform Online. *10,* 197-204.

Kamvar, Z.N., Tabima, J.F., and Grünwald, N.J. (2014). Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. PeerJ *2*, e281.

Karell, P., Ahola, K., Karstinen, T., Valkama, J., and Brommer, J.E. (2011). Climate change drives microevolution in a wild bird. Nat. Commun. *2*, 208.

Kemp, T.S. (2007). The origin of higher taxa: macroevolutionary processes, and the case of the mammals. Acta Zool. *88*, 3–22.

Keogh, J.S., Scott, I.A.W., and Hayes, C. (2005). Rapid and repeated origin of insular gigantism and dwarfism in australian tiger snakes. Evolution *59*, 226–233.

Kern, A.D., and Hahn, M.W. (2018). The Neutral Theory in Light of Natural Selection. Mol. Biol. Evol. *35*, 1366–1371.

Kimura, M. (1957). Some problems of stochastic processes in genetics. The Annals of Mathematical Statistics *28*, 882-901.

Kimura. M. (1960). Optimum mutation rate and degree of dominance as determined by the principle of minimum genetic load. Journal of Genetics *57*, 21-34.

Kimura, M. (1962). On the Probability of Fixation of Mutant Genes in a Population. Genetics *47*, 713–719.

Kimura, M. (1967). On the evolutionary adjustment of spontaneous mutation rates. Genetics Research 9, 23-34.

Kimura, M. (1968). Evolutionary rate at the molecular level. Nature *217*, 624–626.

Kimura, M., Maruyama, T. (1969). The substitutional load in a finite population. Heredity *24*, 101-114

Kimura, M. (1983). The Neutral Theory of Molecular Evolution (Cambridge University Press).

Kimura, M. (1991). The neutral theory of molecular evolution: a review of recent evidence. Idengaku Zasshi *66*, 367–386.

Kimura, M., and Crow, J.F. (1964). The Number of Alleles That Can Be Maintained in a Finite Population. Genetics *49*, 725–738.

Kimura, M., and Ohta, T. (1969). The Average Number of Generations until Fixation of a Mutant Gene in a Finite Population. Genetics *61*, 763–771.

Kimura, M., and Ohta, T. (1971). Protein polymorphism as a phase of molecular evolution. Nature *229*, 467–469.

Kimura, M., Maruyama, T., and Crow, J.F. (1963). The Mutation Load in Small Populations. Genetics *48*, 1303–1312.

King, J.L., and Jukes, T.H. (1969). Non-Darwinian evolution. Science *164*, 788–798.

King, M.C., and Wilson, A.C. (1975). Evolution at two levels in humans and chimpanzees. Science *188*, 107–116.

King, J.L. (1983). This week's citation classic. Current Contents *34*, 25.

Kishida, T. (2017). Population history of Antarctic and common minke whales inferred from individual whole-genome sequences. Mar. Mammal Sci. *33*, 645–652.

Klein, D.R., Meldgaard, M., and Fancy, S.G. (1987). Factors Determining Leg Length in Rangifer tarandus. J. Mammal. *68*, 642–655.

Klicka, J., and Zink, R.M. (1997). The Importance of Recent Ice Ages in Speciation: A Failed Paradigm. Science *277*, 1666–1669.

Klicka, J., and Zink, R.M. (1999). Pleistocene effects on North American songbird evolution. Proc. R. Soc. Lond. B Biol. Sci. *266*, 695–700.

Kosakovsky Pond, S. L., Frost, S. D. W., Muse, S. V. (2005). HyPhy: hypothesis testing using phylogenies. Bioinformatics. *21(5)*, 676-679.

Kosiol, C., Vinar, T., da Fonseca, R.R., Hubisz, M.J., Bustamante, C.D., Nielsen, R., and Siepel, A. (2008). Patterns of positive selection in six Mammalian genomes. PLoS Genet. *4*, e1000144.

Kotlík, P., Marková, S., Konczal, M., Babik, W., and Searle, J.B. (2018). Genomics of end-Pleistocene population replacement in a small mammal. Proc. R. Soc. B Biol. Sci. *285*, 20172624.

Kreitman, M. (1983). Nucleotide polymorphism at the alcohol dehydrogenase locus of Drosophila melanogaster. Nature *304*, 412–417.

Kreitman, M. (1996). The neutral theory is dead. Long live the neutral theory. BioEssays News Rev. Mol. Cell. Dev. Biol. *18*, 678–683; discussion 683.

Kropatsch, R., Dekomien, G., Akkad, D.A., Gerding, W.M., Petrasch-Parwez, E., Young, N.D., Altmüller, J., Nürnberg, P., Gasser, R.B., and Epplen, J.T. (2013). SOX9 duplication linked to intersex in deer. PloS One *8*, e73734.

Kumar, S., and Subramanian, S. (2002). Mutation rates in mammalian genomes. Proc. Natl. Acad. Sci. *99*, 803–808.

Kumar, V., Kutschera, V.E., Nilsson, M.A., and Janke, A. (2015). Genetic signatures of adaptation revealed from transcriptome sequencing of Arctic and red foxes. BMC Genomics *16*, 585.

Laine, V.N., Gossmann, T.I., Schachtschneider, K.M., Garroway, C.J., Madsen, O., Verhoeven, K.J.F., Jager, V. de, Megens, H.-J., Warren, W.C., Minx, P., et al. (2016). Evolutionary signals of selection on cognition from the great tit genome and methylome. Nat. Commun. *7*, ncomms10474.

Laird, C.D., McCONAUGHY, B.L., and McCARTHY, B.J. (1969). Rate of Fixation of Nucleotide Substitutions in Evolution. Nature *224*, 149–154.

Lambeck, K., and Chappell, J. (2001). Sea level change through the last glacial cycle. Science *292*, 679–686.

Lambert, J.W., Reichard, M., and Pincheira-Donoso, D. (2019). Live fast, diversify non-adaptively: evolutionary diversification of exceptionally short-lived annual killifishes. BMC Evol. Biol. *19*, 10.

Lamichhaney, S., Berglund, J., Almén, M.S., Maqbool, K., Grabherr, M., Martinez-Barrio, A., Promerová, M., Rubin, C.-J., Wang, C., Zamani, N., et al. (2015). Evolution of Darwin's finches and their beaks revealed by genome sequencing. Nature *518*, 371–375.

Lamichhaney, S., Han, F., Webster, M.T., Andersson, L., Grant, B.R., and Grant, P.R. (2018). Rapid hybrid speciation in Darwin's finches. Science *359*, 224–228.

Lampert, K.P., Bernal, X.E., Rand, A.S., Mueller, U.G., and Ryan, M.J. (2007). Island Populations of Physalaemus pustulosus: History Influences Genetic Diversity and Morphology. Herpetologica *63*, 311–319.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

Langley, C.H., and Fitch, W.M. (1974). An examination of the constancy of the rate of molecular evolution. J. Mol. Evol. *3*, 161–177.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

Leader-Williams, N. (1980). Population Dynamics and Mortality of Reindeer Introduced into South Georgia. J. Wildl. Manag. *44*, 640–657.

Leader-Williams, N. (1982). Relationship Between a Disease, Host Density and Mortality in a Free-Living Deer Population. J. Anim. Ecol. *51*, 235–240.

Leader-Williams, N. (1988). Reindeer on South Georgia.

Lee, K.M., and Coop, G. (2019). Population genomics perspectives on convergent adaptation. Philos. Trans. R. Soc. Lond. B. Biol. Sci. *374*, 20180236.

Lee, Y.S., Markov, N., Argunov, A., Voloshina, I., Bayarlkhagva, D., Kim, B.-J., Min, M.-S., Lee, H., and Kim, K.S. (2016). Genetic diversity and phylogeography of Siberian roe deer, Caproulus pygargus, in central and peripheral populations. Ecol. Evol. *6*, 7286–7297.

Lepais, O., and Weir, J.T. (2014). SimRAD: an R package for simulation-based prediction of the number of loci expected in RADseq and similar genotyping by sequencing approaches. Mol. Ecol. Resour. *14*, 1314–1321.

Lewontin, R. C., and Hubby, J.L. (1966). A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of Drosophila pseudoobscura. Genetics *54*, 595-609.

Lewontin, R.C. (1974). The Genetic Basis of Evolutionary Change (Columbia University Press).

Lewontin, R.C., and Krakauer, J. (1973). Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics *74*, 175–195.

Li, H., and Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. Nature *475*, 493–496.

Li, C.-Y., Cha, W., Luder, H.-U., Charles, R.-P., McMahon, M., Mitsiadis, T.A., and Klein, O.D. (2012). E-cadherin regulates the behavior and fate of epithelial stem cells and their progeny in the mouse incisor. Dev. Biol. *366*, 357–366.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Li, M., Tian, S., Jin, L., Zhou, G., Li, Y., Zhang, Y., Wang, T., Yeung, C.K.L., Chen, L., Ma, J., et al. (2013). Genomic analyses identify distinct patterns of selection in domesticated pigs and Tibetan wild boars. Nat. Genet. *45*, 1431–1438.

Li, W.H., Wu, C.I., and Luo, C.C. (1985). A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol. Biol. Evol. *2*, 150–174.

Li, Y.F., Costello, J.C., Holloway, A.K., and Hahn, M.W. (2008). "Reverse ecology" and the power of population genomics. Evol. Int. J. Org. Evol. *62*, 2984–2994.

Li, Z., Lin, Z., Ba, H., Chen, L., Yang, Y., Wang, K., Qiu, Q., Wang, W., and Li, G. (2017). Draft genome of the reindeer (Rangifer tarandus). GigaScience *6*, 1–5.

Lischer, H.E.L., and Excoffier, L. (2012). PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. Bioinformatics *28*, 298–299.

Lister, A.M. (1989). Rapid dwarfing of red deer on Jersey in the last interglacial. Nature *342*, 539–542.

Lister, A.M. (2004). The impact of Quaternary Ice Ages on mammalian evolution. Philos. Trans. R. Soc. B Biol. Sci. *359*, 221–241.

Liu, X., and Fu, Y.-X. (2015). Exploring Population Size Changes Using SNP Frequency Spectra. Nat. Genet. *47*, 555–559.

Liu, S., Lorenzen, E.D., Fumagalli, M., Li, B., Harris, K., Xiong, Z., Zhou, L., Korneliussen, T.S., Somel, M., Babbitt, C., et al. (2014). Population genomics reveal recent speciation and rapid evolutionary adaptation in polar bears. Cell *157*, 785–794.

Liu, Y., Rossiter, S.J., Han, X., Cotton, J.A., and Zhang, S. (2010). Cetaceans on a molecular fast track to ultrasonic hearing. Curr. Biol. CB *20*, 1834–1839.

Locke, D.P., Hillier, L.W., Warren, W.C., Worley, K.C., Nazareth, L.V., Muzny, D.M., Yang, S.-P., Wang, Z., Chinwalla, A.T., Minx, P., et al. (2011). Comparative and demographic analysis of orangutan genomes. Nature *469*, 529–533.

Lohmueller, K.E., Albrechtsen, A., Li, Y., Kim, S.Y., Korneliussen, T., Vinckenbosch, N., Tian, G., Huerta-Sanchez, E., Feder, A.F., Grarup, N., et al. (2011). Natural selection affects multiple aspects of genetic variation at putatively neutral sites across the human genome. PLoS Genet. *7*, e1002326.

Lomolino, M.V., van der Geer, A.A., Lyras, G.A., Palombo, M.R., Sax, D.F., and Rozzi, R. (2013). Of mice and mammoths: generality and antiquity of the island rule. J. Biogeogr. *40*, 1427–1439.

Losos, J.B., and Ricklefs, R.E. (2009). Adaptation and diversification on islands. Nature *457*, 830–836.

Lotterhos, K.E., and Whitlock, M.C. (2014). Evaluation of demographic history and neutral parameterization on the performance of FST outlier tests. Mol. Ecol. *23*, 2178–2192.

Lotterhos, K.E., and Whitlock, M.C. (2015). The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. Mol. Ecol. *24*, 1031–1046.

Lourenço, A., Sequeira, F., Buckley, D., and Velo-Antón, G. (2018). Role of colonization history and species-specific traits on contemporary genetic variation of two salamander species in a Holocene island-mainland system. J. Biogeogr. *45*, 1054–1066.

Lovatt, F.M., and Hoelzel, A.R. (2011). The impact of population bottlenecks on fluctuating asymmetry and morphological variance in two separate populations of reindeer on the island of South Georgia. Biol. J. Linn. Soc. *102*, 798–811.

Lovatt, F.M., and Hoelzel, A.R. (2014). Impact on Reindeer (Rangifer tarandus) Genetic Diversity from Two Parallel Population Bottlenecks Founded from a Common Source. Evol. Biol. *41*, 240–250.

Luu, K., Bazin, E., and Blum, M.G.B. (2017). pcadapt: an R package to perform genome scans for selection based on principal component analysis. Mol. Ecol. Resour. *17*, 67–77.

MacArthur, R.H., and Wilson, E.O. (2001). The Theory of Island Biogeography (Princeton University Press).

Malinsky, M., Svardal, H., Tyers, A.M., Miska, E.A., Genner, M.J., Turner, G.F., and Durbin, R. (2018). Whole-genome sequences of Malawi cichlids reveal multiple radiations interconnected by gene flow. Nat. Ecol. Evol. *2*, 1940–1955.

Mallick, S., Gnerre, S., Muller, P., Reich, D. (2009). The difficulty of avoiding false positives in genome scans for natural selection. Genome Res. *19(5): 922-33*.

Margoliash, E. (1963). PRIMARY STRUCTURE AND EVOLUTION OF CYTOCHROME C. Proc. Natl. Acad. Sci. U. S. A. *50*, 672–679.

Martin, S.H., Dasmahapatra, K.K., Nadeau, N.J., Salazar, C., Walters, J.R., Simpson, F., Blaxter, M., Manica, A., Mallet, J., and Jiggins, C.D. (2013). Genome-wide evidence for speciation with gene flow in Heliconius butterflies. Genome Res. *23*, 1817–1828.

Mattle-Greminger, M.P., Bilgin Sonay, T., Nater, A., Pybus, M., Desai, T., de Valles, G., Casals, F., Scally, A., Bertranpetit, J., Marques-Bonet, T., et al. (2018). Genomes reveal marked differences in the adaptive evolution between orangutan species. Genome Biol. *19*, 193.

Maynard-Smith, J. (1968). "Haldane's dilemma" and the rate of evolution. Nature *219*, 1114–1116.

Maynard-Smith, J., and Haigh, J. (1974). The hitch-hiking effect of a favourable gene. Genet. Res. *23*, 23–35.

Mayr, E. (1954). Change of genetic environment and evolution.

Mayr, E. (1999). Systematics and the Origin of Species, from the Viewpoint of a Zoologist (Harvard University Press).

McDonald, J., and Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in Drosophila. Nature.

McFadden, K.W., Gompper, M.E., Valenzuela, D.G., and Morales, J.C. (2008). Evolutionary history of the critically endangered Cozumel dwarf carnivores inferred from mitochondrial DNA analyses. J. Zool. *276*, 176–186.

McKinnon, J.S., Mori, S., Blackman, B.K., David, L., Kingsley, D.M., Jamieson, L., Chou, J., and Schluter, D. (2004). Evidence for ecology's role in speciation. Nature *429*, 294–298.

McVean, G.A.T., Cardin, N.J. (2005). Approximating the coalescent with recombination. Philos. Trans. R. Soc. Lond. B. Biol. Sci. *360*, 1387-1393.

Messier, W. and Stewart, C.B. (1997). Episodic adaptive evolution of primate lysozymes. Nature. *385*, 151-154.

Meynert, A. M., Ansari, M., FitzPatrick, D. R., and Taylor, M. S. (2014). Variant detection sensitivity and biases in whole genome and exome sequencing. BMC Bioinformatics. *15(1).*

Milano, I., Babbucci, M., Cariani, A., Atanassova, M., Bekkevold, D., Carvalho, G.R., Espiñeira, M., Fiorentino, F., Garofalo, G., Geffen, A.J., et al. (2014). Outlier SNP markers reveal fine-scale genetic structuring across European hake populations (Merluccius merluccius). Mol. Ecol. *23*, 118–135.

Momigliano, P., Jokinen, H., Fraimout, A., Florin, A.-B., Norkko, A., and Merilä, J. (2017). Extraordinarily rapid speciation in a marine fish. Proc. Natl. Acad. Sci. U. S. A. *114*, 6074–6079.

Montesinos, D., Santiago, G., and Callaway, R.M. (2012). Neo-allopatry and rapid reproductive isolation. Am. Nat. *180*, 529–533.

Montgomery, W.I., Provan, J., McCabe, A.M., and Yalden, D.W. (2014). Origin of British and Irish mammals: disparate post-glacial colonisation and species introductions. Quat. Sci. Rev. *98*, 144–165.

Morales, A.E., Jackson, N.D., Dewey, T.A., O'Meara, B.C., and Carstens, B.C. (2017). Speciation with Gene Flow in North American Myotis Bats. Syst. Biol. *66*, 440–452.

Mugal, C.F., Wolf, J.B.W., and Kaj, I. (2014). Why time matters: codon evolution and the temporal dynamics of dN/dS. Mol. Biol. Evol. *31*, 212–231.

Murrell, B., Wertheim, J. O., Moola, S., Weighill, T., Scheffler, K., Kosakovsky Pond, S. L. (2012) Detecting individual sites subject to episodic diversifying selection. Plos Genetics.

Murrell, B., Weaver, S., Smith, M. D., Wertheim, J. O., Murrell, S., Aylward, A., Eren, K., Pollner, T., Martin, D. P., Smith, D. M., Scheffler, K., Kosakovsky Pond, S. L. (2015). Gene-wide identification of episodic selection. Molecular Biology and Evolution. *32(5)*, 1365-1371.

Mussmann, S.M., Douglas, M.R., Chafin, T.K., and Douglas, M.E. (2019). BA3-SNPs: Contemporary migration reconfigured in BayesAss for next-generation sequence data. Methods Ecol. Evol. *10*, 1808–1813.

Nadachowska-Brzyska, K., Li, C., Smeds, L., Zhang, G., and Ellegren, H. (2015). Temporal Dynamics of Avian Populations during Pleistocene Revealed by Whole-Genome Sequences. Curr. Biol. CB *25*, 1375–1380.

Nadachowska-Brzyska, K., Burri, R., Smeds, L., and Ellegren, H. (2016). PSMC analysis of effective population sizes in molecular ecology and its application to black-and-white Ficedula flycatchers. Mol. Ecol. *25*, 1058–1072.

Narasimhan, V., Danecek, P., Scally, A., Xue, Y., Tyler-Smith, C., and Durbin, R. (2016). BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data. Bioinforma. Oxf. Engl. *32*, 1749–1751.

Narum, S.R., and Hess, J.E. (2011). Comparison of FST outlier tests for SNP loci under selection. Mol. Ecol. Resour. *11*, 184–194.

Nei, M. (1972). Genetic Distance between Populations. Am. Nat. *106*, 283–292.

Nei, M., Maruyama, T., and Chakraborty, R. 1975, Bottleneck effect and genetic variability in populations. Evolution. *29*, 1-10.

Nei, M., and Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. *3*, 418–426.

Nei, M., and Li, W.H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. U. S. A. *76*, 5269–5273.

Nei, M., Suzuki, Y., and Nozawa, M. (2010). The neutral theory of molecular evolution in the genomic era. Annu. Rev. Genomics Hum. Genet. *11*, 265–289.

Nielsen, R., and Yang, Z. (1998). Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics *148*, 929–936.

Nielsen, E.S., Henriques, R., Toonen, R.J., Knapp, I.S.S., Guo, B., and von der Heyden, S. (2018). Complex signatures of genomic variation of two non-model marine species in a homogeneous environment. BMC Genomics *19*, 347.

Nilsen, E.B., Gaillard, J.-M., Andersen, R., Odden, J., Delorme, D., van Laere, G., and Linnell, J.D.C. (2009). A slow life in hell or a fast life in heaven: demographic analyses of contrasting roe deer populations. J. Anim. Ecol. *78*, 585–594.

Nozawa, M., Suzuki, Y., and Nei, M. (2009). Reliabilities of identifying positive selection by the branch-site and the site-prediction methods. Proc. Natl. Acad. Sci. U. S. A. *106*, 6700–6705.

Nunney, L. (2003). The cost of natural selection revisited. Ann. Zool. Fenn. *40*, 185–194.

Ohta, T. (1992). The Nearly Neutral Theory of Molecular Evolution. Annu. Rev. Ecol. Syst. *23*, 263–286.

Ohta, T. (1993). An examination of the generation-time effect on molecular evolution. Proc. Natl. Acad. Sci. U. S. A. *90*, 10676–10680.

Ohta, T. (1995). Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory. J. Mol. Evol. *40*, 56–63.

Ohta and Gillespie (1996). Development of Neutral and Nearly Neutral Theories. Theor. Popul. Biol. *49*, 128–142.

Oleksyk, T.K., Zhao, K., De La Vega, F.M., Gilbert, D.A., O'Brien, S.J., and Smith, M.W. (2008). Identifying selected regions from heterozygosity and divergence using a light-coverage genomic dataset from two human populations. PloS One *3*, e1712.

Oleksyk, T.K., Smith, M.W., and O'Brien, S.J. (2010). Genome-wide scans for footprints of natural selection. Philos. Trans. R. Soc. B Biol. Sci. *365*, 185–205.

Orr, H.A., and Orr, L.H. (1996). Waiting for Speciation: The Effect of Population Subdivision on the Time to Speciation. Evolution *50*, 1742–1749.

Otte, D., Endler, J. A. (1989). Speciation and Its Consequences (Sunderland, Massachusetts: Sinauer Associates Inc.,U.S.).

Paetkau, D., Calvert, W., Stirling, I., and Strobeck, C. (1995). Microsatellite analysis of population structure in Canadian polar bears. Mol. Ecol. *4*, 347–354.

Paradis, E., and Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinforma. Oxf. Engl. *35*, 526–528.

Paradis, E., Claude, J., and Strimmer, K. (2004). APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics *20*, 289–290.

Parker, J., Tsagkogeorga, G., Cotton, J.A., Liu, Y., Provero, P., Stupka, E., and Rossiter, S.J. (2013). Genome-wide signatures of convergent evolution in echolocating mammals. Nature *502*, 228–231.

Pastene, L.A., Goto, M., Kanda, N., Zerbini, A.N., Kerem, D., Watanabe, K., Bessho, Y., Hasegawa, M., Nielsen, R., Larsen, F., et al. (2007). Radiation and speciation of pelagic organisms during periods of global warming: the case of the common minke whale, Balaenoptera acutorostrata. Mol. Ecol. *16*, 1481–1495.

Patton, A. H., Margres, M. J., Stahlke, A. R., Hendricks, S., Lewallen, K., Hamede, R. K., Ruiz-Aravena, M., Ryder, O., McCallum, H. I., Jones, M. E., Hohenlohe, P. A., Storfer, A. (2019). Contemporary demographic reconstruction methods are robust to genome assembly quality: a case study in Tasmanian devils. Mol. Biol. Evol. *36(12)*, 2906-2921.

Pembleton, L.W., Cogan, N.O.I., and Forster, J.W. (2013). StAMPP: an R package for calculation of genetic differentiation and structure of mixed-ploidy level populations. Mol. Ecol. Resour. *13*, 946–952.

Pennell, M. W., Harmon, L. J., and Uyeda, J.C. (2013). Is there room for punctuated equilibrium in macroevolution? Trends in Ecolofy and Evolution. *29*, 23-32.

Peterson, B.K., Weber, J.N., Kay, E.H., Fisher, H.S., and Hoekstra, H.E. (2012). Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species. PLOS ONE *7*, e37135.

Petit, R.J., Aguinagalde, I., de Beaulieu, J.-L., Bittkau, C., Brewer, S., Cheddadi, R., Ennos, R., Fineschi, S., Grivet, D., Lascoux, M., et al. (2003). Glacial refugia: hotspots but not melting pots of genetic diversity. Science *300*, 1563–1565.

Plakhina, D.A., Zvychanaya, E.Y., Kholodova, M.V., and Danilkin, A.A. (2014). Identification of European (*Capreolus capreolus* L.) and Siberian (*C. pygargus* Pall.) roe deer hybrids by microsatellite marker analysis. Genetika *50*, 862–867.

Poelstra, J.W., Vijay, N., Bossu, C.M., Lantz, H., Ryll, B., Müller, I., Baglione, V., Unneberg, P., Wikelski, M., Grabherr, M.G., et al. (2014). The genomic landscape underlying phenotypic integrity in the face of gene flow in crows. Science *344*, 1410–1414.

Prado-Martinez, J., Sudmant, P.H., Kidd, J.M., Li, H., Kelley, J.L., Lorente-Galdos, B., Veeramah, K.R., Woerner, A.E., O'Connor, T.D., Santpere, G., et al. (2013). Great ape genetic diversity and population history. Nature *499*, 471–475.

Pritchard, J.K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. Genetics *155*, 945–959.

Prüfer, K., Munch, K., Hellmann, I., Akagi, K., Miller, J.R., Walenz, B., Koren, S., Sutton, G., Kodira, C., Winer, R., et al. (2012). The bonobo genome compared with the chimpanzee and human genomes. Nature *486*, 527–531.

Purfield, D. C., Berry, D. P., McParland, S. and Bradley, D. G. (2012). Runs of homozygosity and population history in cattle. BMC Genetics *13(70)*,

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., et al. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am. J. Hum. Genet. *81*, 559–575.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Raia, P., and Meiri, S. (2006). The island rule in large mammals: paleontology meets ecology. Evolution *60*, 1731–1742.

Randi, E., Pierpaoli, M., and Danilkin, A. (1998). Mitochondrial DNA polymorphism in populations of Siberian and European roe deer (*Capreolus pygargus* and *C. capreolus*). Heredity *80 (Pt 4)*, 429–437.

Randi, E., Alves, P.C., Carranza, J., Milosevic-Zlatanovic, S., Sfougaris, A., and Mucci, N. (2004). Phylogeography of roe deer (*Capreolus capreolus*) populations: the effects of historical genetic subdivisions and recent nonequilibrium dynamics. Mol. Ecol. *13*, 3071–3083.

Reaney, A.M., Saldarriaga-Córdoba, M., and Pincheira-Donoso, D. (2018). Macroevolutionary diversification with limited niche disparity in a species-rich lineage of cold-climate lizards. BMC Evol. Biol. *18*, 16.

Reznick, D.N., and Ghalambor, C.K. (2001). The population ecology of contemporary adaptations: what empirical studies reveal about the conditions that promote adaptive evolution. Genetica *112–113*, 183–198.

Reznick, D.N., and Ricklefs, R.E. (2009). Darwin's bridge between microevolution and macroevolution. Nature *457*, 837–842.

Riesch, R., Plath, M., and Bierbach, D. (2018). Ecology and evolution along environmental gradients. Curr. Zool. *64*, 193–196.

Rinker, D.C., Specian, N.K., Zhao, S., and Gibbons, J.G. (2019). Polar bear evolution is marked by rapid changes in gene copy number in response to dietary shift. Proc. Natl. Acad. Sci. U. S. A. *116*, 13446–13451.

Robinson, J.A., Ortega-Del Vecchyo, D., Fan, Z., Kim, B.Y., vonHoldt, B.M., Marsden, C.D., Lohmueller, K.E., and Wayne, R.K. (2016). Genomic Flatlining in the Endangered Island Fox. Curr. Biol. CB *26*, 1183–1189.

Roesti, M., Gavrilets, S., Hendry, A.P., Salzburger, W., and Berner, D. (2014). The genomic signature of parallel adaptation from shared genetic variation. Mol. Ecol. *23*, 3944–3956.

Rozzi, R., and Lomolino, M.V. (2017). Rapid Dwarfing of an Insular Mammal - The Feral Cattle of Amsterdam Island. Sci. Rep. *7*, 8820.

Rundell, R.J., and Price, T.D. (2009). Adaptive radiation, nonadaptive radiation, ecological speciation and nonecological speciation. Trends Ecol. Evol. *24*, 394–399.

Sabeti, P.C., Schaffner, S.F., Fry, B., Lohmueller, J., Varilly, P., Shamovsky, O., Palma, A., Mikkelsen, T.S., Altshuler, D., and Lander, E.S. (2006). Positive natural selection in the human lineage. Science *312*, 1614–1620.

Sackton, T.B., Grayson, P., Cloutier, A., Hu, Z., Liu, J.S., Wheeler, N.E., Gardner, P.P., Clarke, J.A., Baker, A.J., Clamp, M., et al. (2019). Convergent regulatory evolution and loss of flight in paleognathous birds. Science *364*, 74–78.

Salas, E.M., Bernardi, G., Berumen, M.L., Gaither, M.R., and Rocha, L.A. RADseq analyses reveal concordant Indian Ocean biogeographic and phylogeographic boundaries in the reef fish Dascyllus trimaculatus. R. Soc. Open Sci. *6*, 172413.

Sarich, V.M., Wilson, A.C. (1973). Generation time and genomic evolution in primates. Science. *179*, 1144-1147.

Sanderson, M.J. (1997). A nonparametric approach to estimating divergence times in the absence of rate constancy. Molecular Biology and Evolution. *14*, 1218-31.

Schilthuizen, M. (2002). Frogs Flies & Dandelions: The Making of Species (Oxford University Press M.D.).

Schilthuizen, M. (2018). Darwin comes to town: how the urban jungle drives evolution.

Schluter, D. (2001). Ecology and the origin of species. Trends Ecol. Evol. *16*, 372–380.

Schluter, D. (2009). Evidence for ecological speciation and its alternative. Science *323*, 737–741.

Schluter, D., Marchinko, K.B., Barrett, R.D.H., and Rogers, S.M. (2010). Natural selection and the genetics of adaptation in threespine stickleback. Philos. Trans. R. Soc. B Biol. Sci. *365*, 2479–2486.

Schoener, T.W. (2011). The Newest Synthesis: Understanding the Interplay of Evolutionary and Ecological Dynamics. Science *331*, 426–429.

Schneider, A., Souvorov, A., Sabath, N., Landan, G., Gonnet, G. H., Graur, D. (2009). Estimates of positive Darwinian selection are inflated by errors in sequencing, annotation and alignment. Genome Biol. Evol. *1*, 114-118.

Schrider, D.R., Hourmozdi, J.N., and Hahn, M.W. (2011). Pervasive multinucleotide mutational events in eukaryotes. Curr. Biol. CB *21*, 1051–1054.

Seabury, C.M., Bhattarai, E.K., Taylor, J.F., Viswanathan, G.G., Cooper, S.M., Davis, D.S., Dowd, S.E., Lockwood, M.L., and Seabury, P.M. (2011). Genome-Wide Polymorphism and Comparative Analyses in the White-Tailed Deer (Odocoileus virginianus): A Model for Conservation Genomics. PLoS ONE *6*.

Searle, J.B., Kotlík, P., Rambau, R.V., Marková, S., Herman, J.S., and McDevitt, A.D. (2009). The Celtic fringe of Britain: insights from small mammal phylogeography. Proc. Biol. Sci. *276*, 4287–4294.

Seehausen, O., Butlin, R.K., Keller, I., Wagner, C.E., Boughman, J.W., Hohenlohe, P.A., Peichel, C.L., Saetre, G.-P., Bank, C., Brännström, Å., et al. (2014). Genomics and the origin of species. Nat. Rev. Genet. *15*, 176–192.

Selander, R.K., Yang, S.Y., Lewontin, R.C., and Johnson, W.E. (1970). Genetic Variation in the Horseshoe Crab (Limulus polyphemus), A Phylogenetic "Relic." Evolution *24*, 402–414.

Selvaraj, A., and Prywes, R. (2003). Megakaryoblastic leukemia-1/2, a transcriptional co-activator of serum response factor, is required for skeletal myogenic differentiation. J. Biol. Chem. *278*, 41977–41987.

Shultz, A.J., Baker, A.J., Hill, G.E., Nolan, P.M., and Edwards, S.V. (2016). SNPs across time and space: population genomic signatures of founder events and epizootics in the House Finch (Haemorhous mexicanus). Ecol. Evol. *6*, 7475–7489.

Silliman, K. (2019). Population structure, genetic connectivity, and adaptation in the Olympia oyster (Ostrea lurida) along the west coast of North America. Evol. Appl. *12*, 923–939.

Smith, N.G.C., and Eyre-Walker, A. (2002). Adaptive protein evolution in Drosophila. Nature *415*, 1022–1024.

Smithies, O. (2012). How it all began: a personal history of gel electrophoresis. Methods Mol. Biol. Clifton NJ *869*, 1–21.

Sobel, J. M., Chen, G. F., Watt, L. R., Schemske, D.W. (2010). The biology of speciation. Evolution. *64(2)*, 295-315

Sokolov, ., and Gromov, . (1990). The contemporary ideas on roe deer systematization: Morphological, ethological and hybridolohigical analysis.

Sommer, R.S., and Zachos, F.E. (2009). Fossil evidence and phylogeography of temperate species: "glacial refugia" and post-glacial recolonization. J. Biogeogr. *36*, 2013–2020.

Sommer, R.S., Zachos, F.E., Street, M., Joris, O., and Benecke, N. (2008). Late Quaternary distribution dynamics and phylogeography of the red deer (Cervus elaphus) in Europe. Quat. Sci. Rev. *27*, 714–733.

Sommer, R.S., Fahlke, J.M., Schmölcke, U., Benecke, N., and Zachos, F.E. (2009). Quaternary history of the European roe deer *Capreolus capreolus*. Mammal Rev. *39*, 1–16.

Spence, J.P., Steinrucken, M., Terhorst, J., and Song, Y.S. (2018). Inference of population history using coalescent HMMs: review and outlook. Curr. Opin. Genet. Dev. *53*, 70–76.

Spuhler, J.N. (1948). On the Number of Genes in Man. Science *108*, 279–280.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinforma. Oxf. Engl. *30*, 1312–1313.

Stewart, J.R., Lister, A.M., Barnes, I., and Dalén, L. (2010). Refugia revisited: individualistic responses of species in space and time. Proc. Biol. Sci. *277*, 661–671.

Storz, J.F. (2005). Using genome scans of DNA polymorphism to infer adaptive population divergence. Mol. Ecol. *14*, 671–688.

Stronen, A.V., Jędrzejewska, B., Pertoldi, C., Demontis, D., Randi, E., Niedziałkowska, M., Borowik, T., Sidorovich, V.E., Kusak, J., Kojola, I., et al. (2015). Genome-wide analyses suggest parallel selection for universal traits may eclipse local environmental selection in a highly mobile carnivore. Ecol. Evol. *5*, 4410–4425.

Stuart, A.J. (1995). Insularity and Quaternary vertebrate faunas in Britain and Ireland. Geological Society.

Stuessy, T.F., Jakubowsky, G., Gómez, R.S., Pfosser, M., Schlüter, P.M., Fer, T., Sun, B.-Y., and Kato, H. (2006). Anagenetic evolution in island plants. J. Biogeogr. *33*, 1259–1265.

Sturt, F., Garrow, D., and Bradley, S. (2013). New models of North West European Holocene palaeogeography and inundation. J. Archaeol. Sci. *40*, 3963–3976.

Sved, J.A. (1968). Possible Rates of Gene Substitution in Evolution. Am. Nat. *102*, 283–293.

Swärd, K., Stenkula, K.G., Rippe, C., Alajbegovic, A., Gomez, M.F., and Albinsson, S. (2016). Emerging roles of the myocardin family of proteins in lipid and glucose metabolism. J. Physiol. *594*, 4741–4752.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics *123*, 585–595.

Tataru, P., Mollion, M., Glémin, S., and Bataillon, T. (2017). Inference of Distribution of Fitness Effects and Proportion of Adaptive Substitutions from Polymorphism Data. Genetics *207*, 1103–1119.

Templeton, A.R. (2008). The reality and importance of founder speciation in evolution. BioEssays News Rev. Mol. Cell. Dev. Biol. *30*, 470–479.

Thorne, J. L., Kishino, H., Painter, I.S., (1998). Estimating the rate of evolution of the rate of molecular evolution. Mol. Biol. Evol. *15(12)*, 1647-57.

Thomas, J.A., Welch, J.J., Lanfear, R., and Bromham, L. (2010). A generation time effect on the rate of molecular evolution in invertebrates. Mol. Biol. Evol. *27*, 1173–1180.

Thomas, L., Kennington, W.J., Evans, R.D., Kendrick, G.A., and Stat, M. (2017). Restricted gene flow and local adaptation highlight the vulnerability of high-latitude reefs to rapid environmental change. Glob. Change Biol. *23*, 2197–2205.

Thomson, V.A., Mitchell, K.J., Eberhard, R., Dortch, J., Austin, J.J., and Cooper, A. (2018). Genetic diversity and drivers of dwarfism in extinct island emu populations. Biol. Lett. *14*, 20170617.

Tsagkogeorga, G., McGowen, M.R., Davies, K.T.J., Jarman, S., Polanowski, A., Bertelsen, M.F., and Rossiter, S.J. (2015). A phylogenomic analysis of the role and timing of molecular adaptation in the aquatic transition of cetartiodactyl mammals. R. Soc. Open Sci. *2*, 150156.

Tunstall, T., Kock, R., Vahala, J., Diekhans, M., Fiddes, I., Armstrong, J., Paten, B., Ryder, O.A., and Steiner, C.C. (2018). Evaluating recovery potential of the northern white rhinoceros from cryopreserved somatic cells. Genome Res. *28*, 780–788.

Uyeda, J.C., Hansen, T.F., Arnold, S.J., and Pienaar, J. (2011). The million-year wait for macroevolutionary bursts. PNAS, *108,* 15908-15913.

Van Kolfschoten, T., and Laban, C. (1995). Pleistocene terrestrial mammal faunas from the North Sea.

Van Valen, L. (1963). Haldane's Dilemma, Evolutionary Rates, and Heterosis. Am. Nat. *97*, 185–190.

Van Wyngaarden, M., Snelgrove, P.V.R., DiBacco, C., Hamilton, L.C., Rodríguez-Ezpeleta, N., Jeffery, N.W., Stanley, R.R.E., and Bradbury, I.R. (2016). Identifying patterns of dispersal, connectivity and selection in the sea scallop, Placopecten magellanicus, using RADseq-derived SNPs. Evol. Appl. *10*, 102–117.

Vandepitte, K., de Meyer, T., Helsen, K., van Acker, K., Roldán-Ruiz, I., Mergeay, J., and Honnay, O. (2014). Rapid genetic adaptation precedes the spread of an exotic plant species. Mol. Ecol. *23*, 2157–2164.

Varki, A., and Altheide, T.K. (2005). Comparing the human and chimpanzee genomes: searching for needles in a haystack. Genome Res. *15*, 1746–1758.

Velo-Antón, G., Zamudio, K.R., and Cordero-Rivera, A. (2012). Genetic drift and rapid evolution of viviparity in insular fire salamanders (Salamandra salamandra). Heredity *108*, 410–418.

Venkat, A., Hahn, M.W., and Thornton, J.W. (2018). Multinucleotide mutations cause false inferences of lineage-specific positive selection. Nat. Ecol. Evol. *2*, 1280–1288.

Via, S. (2009). Natural selection in action during speciation. Proc. Natl. Acad. Sci. U. S. A. *106 Suppl 1*, 9939–9946.

Vijay, N., Park, C., Oh, J., Jin, S., Kern, E., Kim, H.W., Zhang, J., and Park, J.-K. (2018). Population Genomic Analysis Reveals Contrasting Demographic Changes of Two Closely Related Dolphin Species in the Last Glacial. Mol. Biol. Evol. *35*, 2026–2033.

Vogel, F. (1964). A preliminary estimate of the number of human genes. Nature *201*, 847.

Vorobieva, N.V., Sherbakov, D.Y., Druzhkova, A.S., Stanyon, R., Tsybankov, A.A., Vasil'ev, S.K., Shunkov, M.V., Trifonov, V.A., and Graphodatsky, A.S. (2011). Genotyping of *Capreolus pygargus* Fossil DNA from Denisova Cave Reveals Phylogenetic Relationships between Ancient and Modern Populations. PLOS ONE *6*, e24045.

Vrba, E.S., and DeGusta, D. (2004). Do species populations really start small? New perspectives from the Late Neogene fossil record of African mammals. Philos. Trans. R. Soc. Lond. B. Biol. Sci. *359*, 285-292; discussion 292-293.

Wagner, A. (2007). Rapid Detection of Positive Selection in Genes and Genomes Through Variation Clusters. Genetics *176*, 2451–2463.

Wang, S., Zhu, W., Gao, X., Li, X., Yan, S., Liu, X., Yang, J., Gao, Z., and Li, Y. (2014). Population size and time since island isolation determine genetic diversity loss in insular frog populations. Mol. Ecol. *23*, 637–648.

Wang, X., Que, P., Heckel, G., Hu, J., Zhang, X., Chiang, C.-Y., Zhang, N., Huang, Q., Liu, S., Martinez, J., et al. (2019). Genetic, phenotypic and ecological differentiation suggests incipient speciation in two Charadrius plovers along the Chinese coast. BMC Evol. Biol. *19*, 135.

Waterson, R., Lander, E., and Wilson, R. (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. Nature *437*, 69–87.

Watterson, G.A. (1975). On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. *7*, 256–276.

Weigand, H., and Leese, F. (2018). Detecting signatures of positive selection in non-model species using genomic data. Zool. J. Linn. Soc. *184*, 528–583.

Weigelt, P., Jetz, W., and Kreft, H. (2013). Bioclimatic and physical characterization of the world's islands. Proc. Natl. Acad. Sci. *110*, 15307–15312.

Weiner, J. (1994). The beak of the finch: a story of evolution in our time. New York Knopf.

Weir, B.S., and Cockerham, C.C. (1984). Estimating F-Statistics for the Analysis of Population Structure. Evolution *38*, 1358–1370.

Weissman, D.B., and Barton, N.H. (2012). Limits to the Rate of Adaptive Substitution in Sexual Populations. PLoS Genet. *8*.

Wellenreuther, M., and Sánchez-Guillén, R.A. (2016). Nonadaptive radiation in damselflies. Evol. Appl. *9*, 103–118.

Wertheim, J. O., Murrell, B., Smith, M. D., Kosakovsky Pond, S. L., Scheffler, K. (2015). RELAX: detecting relaxed selection in a phylogenetic framework. *32(3)*, 820-832.

Wiehler, J., and Tiedemann, R. (1998). Phylogeography of the European roe deer *Capreolus capreolus* as revealed by sequence analysis of the mitochondrial Control Region.

Wiens, J.J. (2004). Speciation and ecology revisited: phylogenetic niche conservatism and the origin of species. Evolution *58*, 193–197.

Willi, Y., van Buskirk, J., and Hoffmann, A.A. (2006). Limits to the Adaptive Potential of Small Populations. Annu. Rev. Ecol. Evol. Syst. *37*, 433–458.

Wilson, A.C., and Sarich, V.M. (1969). A molecular time scale for human evolution. Proc. Natl. Acad. Sci. U. S. A. *63*, 1088–1093.

Winker, K., Glenn, T.C., Withrow, J., Sealy, S.G., and Faircloth, B.C. (2019). Speciation despite gene flow in two owls (Aegolius ssp.): Evidence from 2,517 ultraconserved element loci. The Auk *136*, 1–12.

Wisotsky, S. R., Kosakovsky Pond, S. L., Shank, S. D., and Muse, S. V. (2020). Synonymous site-to-site substitution rate variation dramatically inflates false positive rates of selection analyses: ignore at your own peril. Molecular Biology and Evolution. *37(8)*, 2430-2439.

Wolf, J.B.W., Künstner, A., Nam, K., Jakobsson, M., and Ellegren, H. (2009). Nonlinear dynamics of nonsynonymous (dN) and synonymous (dS) substitution rates affects inference of selection. Genome Biology and Evolution. *1*, 308-319.

Wolf, J.B.W., and Ellegren, H. (2017). Making sense of genomic islands of differentiation in light of speciation. Nat. Rev. Genet. *18*, 87–100.

Wright, S. (1931). Evolution in Mendelian Populations. Genetics *16*, 97–159.

Wu, C.I., and Li, W.H. (1985). Evidence for higher rates of nucleotide substitution in rodents than in man. Proc. Natl. Acad. Sci. U. S. A. *82*, 1741–1745.

Wu, D.-D., Ding, X.-D., Wang, S., Wójcik, J.M., Zhang, Y., Tokarska, M., Li, Y., Wang, M.-S., Faruque, O., Nielsen, R., et al. (2018). Pervasive introgression facilitated domestication and adaptation in the Bos species complex. Nat. Ecol. Evol. *2*, 1139–1145.

Xiao, C.-T., Zhang, M.-H., Fu, Y., and Koh, H.-S. (2007). Mitochondrial DNA distinction of northeastern China roe deer, Siberian roe deer, and European roe deer, to clarify the taxonomic status of northeastern China roe deer. Biochem. Genet. *45*, 93–102.

Xu, H., Liu, Y., He, G., Rossiter, S.J., and Zhang, S. (2013). Adaptive evolution of tight junction protein claudin-14 in echolocating whales. Gene *530*, 208–214.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. *24*, 1586–1591.

Yang, Z., and dos Reis, M. (2011). Statistical properties of the branch-site test of positive selection. Mol. Biol. Evol. *28*, 1217–1228.

Yoder. A.D. and Yang, Z. (2000). Estimation of primate speciation dates using local molecular clocks. Molecular Biology and Evolution. *17*, 1081-1090.

Zeileis, A. (2005). zoo: S3 Infrastructure for Regular and Irregular Time Series. ResearchGate.

Zepeda Mendoza, M.L., Xiong, Z., Escalera-Zamudio, M., Runge, A.K., Thézé, J., Streicker, D., Frank, H.K., Loza-Rubio, E., Liu, S., Ryder, O.A., et al. (2018). Hologenomic adaptations underlying the evolution of sanguivory in the common vampire bat. Nat. Ecol. Evol. *2*, 659–668.

Zhang, L., and Li, W.-H. (2005). Human SNPs reveal no evidence of frequent positive selection. Mol. Biol. Evol. *22*, 2504–2507.

Zhou, T., Enyeart, P.J., and Wilke, C.O. (2008). Detecting Clusters of Mutations. PLoS ONE *3*.

Zhu, L., Deng, C., Zhao, X., Ding, J., Huang, H., Zhu, S., Wang, Z., Qin, S., Ding, Y., Lu, G., et al. (2018). Endangered Père David's deer genome provides insights into population recovering. Evol. Appl. *11*, 2040–2053.

Zimin, A.V., Delcher, A.L., Florea, L., Kelley, D.R., Schatz, M.C., Puiu, D., Hanrahan, F., Pertea, G., Van Tassell, C.P., Sonstegard, T.S., et al. (2009). A whole-genome assembly of the domestic cow, Bos taurus. Genome Biol. *10*, R42.

Zucchi, M.I., Cordeiro, E.M.G., Allen, C., Novello, M., Viana, J.P.G., Brown, P.J., Manjunatha, S., Omoto, C., Pinheiro, J.B., and Clough, S.J. (2019). Patterns of Genome-Wide Variation, Population Differentiation and SNP Discovery of the Red Banded Stink Bug (Piezodorus guildinii). Sci. Rep. *9*, 1–11.

Zuckerkandl, E., and Pauling, L. (1965). Evolutionary Divergence and Convergence in Proteins. Evol. Genes Proteins 97–166.